



BANK OF ENGLAND

# Staff Working Paper No. 688

## Sending firm messages: text mining letters from PRA supervisors to banks and building societies they regulate

David Bholat, James Brookes, Chris Cai, Katy Grundy and Jakob Lund

October 2017

Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee or Prudential Regulation Committee.



BANK OF ENGLAND

# Staff Working Paper No. 688

## Sending firm messages: text mining letters from PRA supervisors to banks and building societies they regulate

David Bholat,<sup>(1)</sup> James Brookes,<sup>(2)</sup> Chris Cai,<sup>(3)</sup> Katy Grundy<sup>(4)</sup> and Jakob Lund<sup>(5)</sup>

### Abstract

Our paper analyses confidential letters sent from the Bank of England's Prudential Regulation Authority (PRA) to banks and building societies it supervises. These letters are a 'report card' written to firms annually, and are arguably the most important, regularly recurring written communication sent from the PRA to firms it supervises. Using a mix of methods, including a machine learning algorithm called random forests, we explore whether the letters vary depending on the riskiness of the firm to whom the PRA is writing. We find that they do. We also look across the letters as a whole to draw out key topical trends and confirm that topics important on the post-crisis regulatory agenda such as liquidity and resolution appear frequently. And we look at how PRA letters differ from the letters written by the PRA's predecessor, the Financial Services Authority. We find evidence that PRA letters are different, with a greater abundance of forward-looking language and directiveness, reflecting the shift in supervisory approach that has occurred in the United Kingdom following the financial crisis of 2007–09.

**Key words:** Bank of England Prudential Regulation Authority, banking supervision, text mining, machine learning, random forests, Financial Services Authority, central bank communications.

**JEL classification:** C55, C80, E58, G28.

---

(1) Bank of England. Email: david.bholat@bankofengland.co.uk

(2) Bank of England. Email: james.brookes@bankofengland.co.uk

(3) Bank of England. Email: chris.cai@bankofengland.co.uk

(4) Bank of England. Email: katy.grundy@bankofengland.co.uk

(5) Bank of England. Email: jakob.lund@bankofengland.co.uk

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 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. 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 Sadia Arif, Alex Holmes and Simon Morley for vetting the paper, Alex Clark and Matthew Gill for originally proposing this research, Arthur Turrell for his suggestion of the title and to Angelina Carvalho, Matthew Everitt and Pedro Santos for their early contributions.

Information on the Bank's working paper series can be found at [www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx](http://www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx)

Publications and Design Team, Bank of England, Threadneedle Street, London, EC2R 8AH  
Telephone +44 (0)20 7601 4030 email [publications@bankofengland.co.uk](mailto:publications@bankofengland.co.uk)

© Bank of England 2017  
ISSN 1749-9135 (on-line)

## 1. Introduction

A decade ago, the global financial crisis manifested itself in the UK with the run on Northern Rock (Bholat and Gray 2013). While not the main cause, inadequate bank supervision played a role in the firm's failure. In a post-mortem internal audit, the Financial Services Authority (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 2008: 23). On other occasions, some messages that should have been communicated were not; for example, about the inadequacy of its stress testing process (Ibid: 32). Overall, the audit found that the FSA had not been sufficiently clear in its official communications with the firm. A Treasury Select Committee report corroborated this conclusion. The report criticised the FSA for communicating too infrequently. When it did communicate, the Committee concluded it did not do so with a clear focus on outcomes (House of Commons Treasury Select Committee 2008: 24). The Treasury Select Committee report concluded that supervisory communication with the firm did not effectively focus on low probability but high impact events that would present risks to the FSA's objectives (Ibid: 30). Similar issues in supervisory communications played themselves out in the lead up to crises at HBOS and Royal Bank of Scotland (FCA and PRA 2015: 14; Financial Services Authority 2011: 258). In part as a result of these failings, the FSA was disbanded and prudential banking supervision handed over to the Prudential Regulation Authority (PRA), part of the Bank of England.

So has supervisory communication with firms improved post-crisis, and how does it vary depending on the firm? To answer this question, we have text mined confidential Periodic Summary Meeting (PSM) letters sent to banks and building societies annually by the PRA. 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 potential impact of the firm and its proximity to resolution. 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 FSA.

Our paper unfolds as follows. The next section of the paper discusses the PRA's overall approach to supervision, and the role PSM letters play within it. The PSM process is the key point in the supervisory year for PRA regulated firms and their supervisors. At the PSM meeting, a panel composed of PRA management and senior advisors assess a firm's risks and the PRA's supervisory strategy related to that firm. After it, a letter is drafted to communicate key messages to the firm. In sum, the PSM letter sets out a summary of the PRA's view on the most material risks facing the firm, the most material risks that firm poses to the PRA's objectives to ensure financial system safety and soundness, and delineates required mitigating actions. The PSM letter is arguably the most important formal communication from the PRA to a firm in the course of the year.<sup>1</sup>

The third section of the paper describes our text mining methodology. Text mining involves the quantification of qualitative data. It refers to a family of computationally-based approaches that use algorithms to find patterns in texts that human readers may be unable to detect. It is an increasingly popular set of methods used by researchers to investigate central

---

<sup>1</sup> This is especially the case for smaller firms because they tend to receive relatively fewer communications from the PRA than larger firms. Other forms of communication to all firms include telephone conversations, regular conference calls, ad-hoc emails, midpoint review updates, authorisation notifications, consultation papers and supervisory statements.

bank communications (Goldsmith-Pinkham, Hirtle and Lucca 2016; Bholat et al 2015; Hansen, McMahon and Prat 2014). A common approach in this literature is to use topic models to surface discursive content. Our paper makes a methodological contribution to the literature by deploying a machine learning algorithm called random forests to measure deep linguistic structure. Furthermore, we develop and measure a set of 25 linguistic features that can be used by future researchers to investigate other central banking texts. While measures of linguistic complexity and sentiment are stock-in-trade for text miners, our paper goes beyond them to quantify directiveness, formality and forward-lookingness as well. In addition, while previous studies have concentrated largely on structural complexity (e.g. sentence length, document length), we have included measures of cognitive complexity that capture facets of language that increase processing burden on readers (e.g. the rate of numerals and acronyms).

In the fourth section, we assess whether and to what extent the PRA's communications with firms are commensurate with the degree of risk they pose to the PRA's statutory objectives—whether those risks stem from a firm's potential impact (inherent risk) or its proximity to resolution (imminent risk). For students of political economy, our paper provides much needed empirical insight into the relationship between banks and their supervisors. Too often, the academic debate on banking supervision is based solely on theoretical priors. In one camp are those concerned with regulatory capture—those who believe the relationship between supervisors and the firms they regulate is *inevitably* too cosy (Kane 2015). In another camp are those that see the relationship as *intrinsically* antagonistic. We find support for neither view based on the tone of the letters.<sup>2</sup> Our sentiment analysis, based on a finance-specific dictionary, indicates that the tone of PSM letters is neutral and professional.

In the fifth section, we explore how the PRA's PSM letters differ from the FSA's ARROW letters. We find that they are very different. In particular, we find that PSM letters are more directive, with a greater abundance of obligative phrases (e.g. *must*, *should*, *expect*) and deadlines. We conclude the paper by drawing out the implications of our findings for PRA supervisors and the general public.

## 2. Background

### The post-crisis creation of the PRA

Following the financial crisis, the FSA was disbanded and replaced by a new 'twin peaks' approach to financial regulation in the UK. The Prudential Regulation Authority (PRA), part of the Bank of England, is responsible for the prudential regulation of UK banks, building societies, insurers, credit unions, and the UK subsidiaries of foreign firms including large investment banks. 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. Conduct regulation of firms is now undertaken by the Financial Conduct Authority (FCA). Its objectives are to secure an appropriate degree of protection for consumers, to enhance the integrity of the UK financial system via regulating markets, and to promote effective competition in the interests of consumers.

---

<sup>2</sup> We wish to stress, however, that we make no claims about the content.

This split formally happened at Legal Cut-over on 1 April 2013.<sup>3</sup> That same year the PRA published its *Approach to Banking Supervision* document, updated in 2016 (Bank of England 2016). The document sets out how the PRA approaches the prudential supervision of deposit takers. Key tenets of the approach include:

- Within the statutory framework, the PRA’s approach relies significantly on supervisory judgement
- The PRA supervises firms to judge whether they are safe and sound, and whether they meet – and are likely to continue to meet – Threshold Conditions<sup>4</sup>
- The PRA’s approach is forward-looking. It not only assesses firms’ current risks, but also those that could plausibly arise in the future
- The PRA focuses on those issues and those firms that pose the greatest risks to the stability of the UK financial system (proportionality)
- The PRA’s regulatory decision-making is rigorous and well-documented, consistent with public statutes and the PRA’s Fundamental Rules<sup>5</sup>

At the heart of the PRA’s approach to supervision is a Risk Model, which supervisors use as a framework for assessing risks posed by firms to PRA objectives. The Risk Model has two high-level aspects. The first is Gross Risk, which comprises the Potential Impact a firm’s failure would have on the financial system; macroeconomic and other risks to which the firm is exposed (External Context); and risks inherent in the firm’s business model and corporate structure (Business Risk). The second aspect of the Risk Model is the Mitigating Factors that offset these risks, including Management and Governance, Risk Management and Controls, Capital, Liquidity, and Resolvability.<sup>6</sup> Figure 1 shows the Risk Model.

---

<sup>3</sup> The actual change was more gradual. In 2011, the FSA re-organised itself into an ‘internal twin peaks’ model of a Conduct Business Unit (CBU) and a Prudential Business Unit (PBU)—embryonic forms of the future FCA and PRA, respectively. By mid-2012, FSA resources already had been almost entirely allocated to the CBU and PBU, with minimal levels of central resourcing. PBU staff physically migrated to new premises near the Bank in phases during the first quarter of 2013, and moved to Bank IT systems at the same time. The development of the PRA also continued beyond Legal Cut-over, especially following the launch of the Bank of England’s 2014 “One Bank” Strategic Plan. The Strategic Plan included an initiative “Delivering Supervision as One Bank” that aimed at enriching micro-prudential supervision with analytical perspectives from the wider Bank, e.g. macro-prudential analysis and collaboration with the Bank’s Special Resolution Directorate on firms’ resolvability (Bank of England 2014).

<sup>4</sup> The Threshold Conditions are a set of minimum requirements that firms must meet in order to be permitted to carry on regulated activities, as defined in the Financial Services and Markets Act 2000 (FSMA). Broadly, they require firms to have an appropriate quantity and quality of capital and liquidity, to have appropriate resources to measure, monitor and manage risk, to be fit and proper, to conduct their business prudently and to be capable of being effectively supervised by the PRA (PRA and FCA 2016). Threshold conditions are assessed at least annually at the PSM meeting, and on an ad-hoc basis in response to material market developments.

<sup>5</sup> The Fundamental Rules set out the PRA’s expectations of firms. These are (1) A firm must conduct its business with integrity; (2) A firm must conduct its business with due skill, care and diligence; (3) A firm must act in a prudent manner; (4) A firm must at all times maintain adequate financial resources; (5) A firm must have effective risk strategies and risk management systems; (6) A firm must organise and control its affairs responsibly and effectively; (7) A firm must deal with the PRA in an open and co-operative way, and must disclose to the PRA appropriately anything relating to the firm of which the PRA would reasonably expect notice; (8) A firm must prepare for resolution so, if need arises, it can be resolved in an orderly manner with a minimum disruption of critical services.

<sup>6</sup> Management and Governance refers to aspects such as the competence of a firm’s senior management, the constitution of its Board, and a firm’s culture. Risk Management and Control includes an assessment of the firm’s risk identification and mitigation processes including operational risk.

Gross risk		Mitigating factors					
Potential impact	Risk context	Operational mitigation		Financial mitigation		Structural mitigation	
Potential impact	External context	Business risk	Management and governance	Risk management and controls	Capital	Liquidity	Resolvability

Figure 1: The PRA's Risk Model. The red box highlights factors that contribute to PIF stage

With the exception of Potential Impact, supervisors use a ten point scale to score a firm along each of these risk elements. A score of 1 indicates the lowest risk to safety and soundness, and 10 the highest. Among other uses, these risk element scores are combined to determine an overall Proactive Intervention Framework (PIF) stage for each firm.<sup>7</sup> The summary PIF stage can be interpreted as a ‘distance to default’ or ‘proximity to failure’ measure. PIF stages run from 1 to 5, with 1 signifying low risks to the viability of the firm, and 5 a firm that is in resolution or being actively wound down. While PIF staging takes into account a firm’s External Context, Business Risk, Management and Governance, Risk Management and Controls, Capital, and Liquidity risk element scores, supervisors use judgement when deciding the weight applied to each. In other words, the PIF stage is not simply a summation and average of the risk element scores, but is the product of a more complex deliberation, reflecting the PRA’s emphasis on judgement in supervision. Furthermore, the PIF stage does not take into account the Potential Impact or Resolvability scores. The Resolvability score considers how easy it would be to resolve a firm in an orderly manner should it fail. As such, it is not relevant to how close a firm is to failure. The Potential Impact score is determined as part of a separate process discussed next. Figure 2 summarises the PIF stages.

Increasing risk 	<b>Stage 1</b>	<b><u>Low risk</u> to viability of firm</b>
	<b>Stage 2</b>	<b><u>Moderate risk</u> to viability of firm</b>
	<b>Stage 3</b>	<b><u>Risk to viability</u> absent action by the firm</b>
	<b>Stage 4</b>	<b><u>Imminent risk</u> to viability of firm</b>
	<b>Stage 5</b>	<b>Firms <u>in resolution</u> or being actively wound up</b>

Figure 2: PIF Stages

<sup>7</sup> A firm’s PIF stage and their risk element scores are never disclosed to them, to avoid risk of public disclosure.

## Firm categorisation and potential impact

While all firms are assessed by supervisors according to the supervisory Risk Model, the frequency, depth and content of that assessment is calibrated according to their Potential Impact categorisation. A firm's category is determined by its Potential Impact score, measured by the extent and scale of a firm's core economic functions e.g. retail banking, payment services, general insurance etc. Each economic function is measured with different data, which are weighted and aggregated, first for each economic function, and then across all core economic functions to calculate an initial Potential Impact score. Supervisors can adjust the Potential Impact score up or down based on their judgement of qualitative factors not captured by the data. Once done, the PRA buckets 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. Accordingly, the PRA supervises Category 1 firms more intensively than lower category firms. A Potential Impact assessment is run annually for all firms by a central team, and the results are then reviewed by a senior PRA executive committee to agree firms' potential impact scores. Scores may be updated sooner in the event of a material change that impacts the firm, such as a merger, acquisition or disposal. Figure 3 summarises the different firm categories.



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

Figure 3: Firm Categories

## The role of the PSM and the PSM letter in supervision

Supervision involves ongoing 'continuous assessment', with PRA supervisors continuously reflecting on whether their supervisory strategy and work-plan for each firm remain appropriate. The most important moment for such reflections is the Periodic Summary Meeting (PSM).<sup>8</sup> These are annual meetings where supervisors explain the key risks posed by the firm to the PRA's objectives; look back at supervisory work conducted over the past twelve months; approve the proposed supervisory plan for the next twelve months; and

<sup>8</sup> Supervisors also pause and reflect every six months to conduct so-called Mid-Point reviews.

reassess the longer term supervisory strategy for the firm. As part of this, the PSM will consider and confirm the firm's categorisation; its risk element score and PIF stage; compliance with Threshold Conditions; the fitness and propriety of senior managers; and, as appropriate, the adequacy of the capital and liquidity resources of the firm.

PSM meetings are a key part of the PRA's decision making process. Each PSM meeting will involve frontline supervisors presenting to a PSM panel including independent senior managers and senior advisors whose role is to provide feedback and challenge on the proposed supervisory strategy and messages. 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 PRA's Board.<sup>9</sup> While the seniority of attendees at PSM meetings will differ for different category firms, the PSM panel always needs an 'independent' member not involved in the day-to-day supervision of the firm. For Category 1, 2 and 3 firms, PSMs are firm-specific meetings. For Category 4 and 5 firms, the PSM may consider groups of firms in the same category together. Even then, each firm is considered on a case-by-case basis.

The outcomes of the PSM meetings are communicated to firms via a PSM letter sent afterward. Broadly speaking, the PSM letter is intended to convey the PRA's judgement 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.

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 PIF stage, and for the highest risk firms additional supervisory communication could potentially be of similar importance to the PSM letter.

### **3. Analysing PSM letters**

The PRA's supervisory approach emphasises proportionality. For example, Category 1 firms are supervised more intensively than Category 5 firms. Similarly, PIF Stage 4 firms will receive more attention than PIF Stage 1 firms because they are closer to resolution. We would therefore expect the letters written to different Category and PIF stage firms to be linguistically different. In the fourth section of this paper, we test that hypothesis. In the fifth section, we then compare PRA PSM letters to FSA ARROW letters to understand how, if at all, supervisory communication has changed since the crisis. First, however, we discuss our data, linguistic measures, and machine learning methodology.

#### **Data, measures and method**

We focused our analysis on a representative sample of comparable UK banks and building societies supervised by the UK Deposit Takers Directorate, with two years of PSM letters amenable to text analysis – 2014 and 2015.<sup>10</sup> Note that as the data are sensitive, we cannot reveal the population or sample size, nor can we reveal the number of observations in each of

---

<sup>9</sup> The PRA's Board was reconstituted in 2017 into a new statutory Prudential Regulation Committee (PRC) of the Bank of England.

<sup>10</sup> We sampled on firm type (bank vs. building society) and category, allowing PIF to vary.

the Category and PIF stages. Details of the data selection criteria and exclusions are included in the first annex. In order to test the out-of-sample robustness of our machine learning models, we used the 2016 letters to the same firms. We also 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.<sup>11</sup>

For all the letters, we constructed a set of 25 linguistic features. These linguistic features are summarised below. Their detailed definition and measurement is described in Annex 2. They can be grouped into five high-level groups:

### 1) Measures of linguistic complexity

We might expect that, if the PRA is being proportionate, then the letters to the greatest Potential Impact firms, which pose greater inherent risk to the PRA’s objectives, and those to firms at higher PIF stages, where the imminent risk is higher, would be more detailed, lengthier, and more complex—for instance, because more specific detail is needed to elucidate the firms’ risks in these cases. In addition, if the PRA’s supervisory approach is more thorough than the FSA’s, we’d also expect to find that PSM letters are more complex than ARROW letters. To explore this, we considered nine complexity features distributed across two sub-types of complexity – that at the document level and that at the sentence level. These are given in Figure 4 below.

Description of Feature
length of letter (in words)
number of section headings in letter <sup>12</sup>
presence of an appendix
proportion of acronyms in letter (out of total number of words)
proportion of numerals in letter (out of total number of words)
mean sentence length (in words)
mean rate of punctuation per sentence
mean rate of subordination per sentence
mean rate of verbs per sentence

Figure 4: Complexity features

### 2) Sentiment indicators

We might expect that firms at higher PIF stages receive more negatively worded letters from the PRA. Similarly, we might expect those firms that pose a larger inherent risk to the PRA’s objectives (firm category) receive more negatively worded letters, as the PRA is likely to be concerned about such firms simply because of the impact they can have on the wider

<sup>11</sup> When comparing the ARROW letters with the PSM letters, we focused on the latest available vintage of PSM letters from 2015 for two reasons. First, since 2014 was the first year in which PSM letters were sent to firms, it is possible that changes which took place in the supervisory approach had not fully permeated supervisory communication. Second, the use of PSM data from a single year provides roughly the same number of observations as the ARROW data, to yield a roughly like-for-like comparison that is balanced overall.

<sup>12</sup> However, it could be alternatively argued that a greater number of section headings may make the text easier to read because they help structure a document into easily digestible chunks. For us, however, it is a proxy for the number of topics mentioned in a letter.

financial system. We had no prior about differences in sentiment expressed by PSM versus ARROW letters. Figure 5 gives the features.<sup>13</sup>

Description of Feature
financial sentiment score
proportion of high risk vocabulary in letter (out of total number of words)

Figure 5: Sentiment features

### 3) Directiveness

We might expect to see more directive language (orders and requests) and direct language ('impoliteness') used with firms where the inherent risks are greater (e.g. Category 1 firms) or where imminent risks are larger (e.g. PIF 4 firms). We might also suppose that the PRA is more assertive in its communication than was the FSA, born as it was after the financial crisis. Figure 6 lists the individual linguistic features we used to measure directiveness.

Description of Feature
proportion of obligative words in letter (out of total number of words)
proportion of deadlines in a letter (out of total number of words)
proportion of 'please' in a letter (out of total number of words)
ratio of sentence-initial 'please' count to sentence-medial 'please' count
ratio of sentence-initial 'you' count to sentence-medial 'you' count

Figure 6: Directiveness features

### 4) Formality

We also explored whether various formality attributes might help to distinguish the letter types, although we had no priors here. Our formality features are listed in Figure 7.

Description of Feature
proportion of local person pronouns in letter (out of total number of words)
ratio of 'I' count to 'PRA' count
ratio of 'I' count to 'we' count
ratio of 'we' count to 'PRA' count
ratio of 'you' count to firm count
whether the salutation is handwritten or not
whether the salutation is to a named individual or not

Figure 7: Formality features

<sup>13</sup> Note that the 'financial sentiment score' refers to the difference between the number of words in the document that express positive sentiment in a financial setting, and the number of words in the document that express negative sentiment in a financial setting, divided by the total number of words in the document. 'High risk vocabulary' refers to words such as *weak*, *vulnerable*, *exposed*, etc. For further details, see the annex.

### 5) Forward-lookingness

A fair prior is that, if the PRA is indeed forward-looking, much of the text in letters will relate to the future. By contrast, the FSA letters might be relatively more backward-looking. The two variables by which we measure forward-lookingness are given in Figure 8.

Description of Feature
proportion of non-past tensed verbs (out of all tense marked verbs)
proportion of future-oriented sentences (out of total number of sentences)

Figure 8: Forward-looking features

In addition, when exploring differences along the PIF and Category classes, we included two non-linguistic features– the firm type (bank vs. building society) and the year of the letter (2014 and 2015).<sup>14,15</sup> Our reason for including these is that they may interact with the linguistic features, and thereby increase discriminative power.

To relate our linguistic features with our response variables of interest – PIF scores, firm category and letter type (ARROW versus PSM) – we used a machine learning algorithm known as random forests. For several reasons, the nature of our data made us choose random forests over other, more familiar statistical techniques such as classical logistic regression. First, regressions are only suitable for “tall” datasets, where the number of observations of the smallest class exceeds the number of features pertaining to them. To visualise this in database terms, regressions are typically valid when there are more rows for the smallest class of observations than columns. However, in our research, some of our class observations (sub-sample of letters) had fewer observations than features. Random forests have been shown to be reliable when working with such “wide” datasets, that is, when the number of features outnumbers the number of observations (see Strobl et al 2009).

Second, while standard regressions try to fit data to a linear form, random forests are capable of indicating complex, non-linear relationships and interactions. Language exhibits exactly these properties. Language is a ‘complex’ system in the complexity science sense that the meaning of a word is not given intrinsically, but arises through its relation to other words (Saussure 1983). In particular, written language exhibits non-linearities because discourse is sensitive to so-called ‘butterfly effects’: small, subtle changes in wording or syntax may result in dramatic shifts in overall meaning.<sup>16</sup> While these nuances of language are difficult to capture using any purely quantitative approach, they are better modelled by random forests than in a regression framework, where non-linearities and interactions have to be pre-specified in the model, based on prior domain-specific knowledge.<sup>17</sup>

---

<sup>14</sup> We also explored models in which the alternative operationalization of risk was included as a feature, i.e. PIF as a predictor of Category and Category as a predictor of PIF. However, we did not find that the relative importance of the top-ranking linguistic features changed dramatically.

<sup>15</sup> It was not sensible to include a feature for the author of the letter, as PSM letters are written by a team of authors and no single author can be identified.

<sup>16</sup> Negation is an obvious example. The sentence "The firm is not in trouble" expresses the diametrically opposite view of the sentence "The firm is in trouble." Similarly, small changes in punctuation can cause big shifts in meaning, as Steven Pinker points out with a humorous example. The phrase "Rachael Ray finds inspiration in cooking, her family, and her dog" has a completely different meaning when the commas are removed: "Rachael Ray finds inspiration in cooking her family and her dog" (Pinker 2014: 121).

<sup>17</sup> Furthermore, the random forest algorithm can sift through a potentially vast range of features, and identify which of them are most strongly related to the response variable via an in-built variable importance procedure which determines the extent to which the accuracy of the model decreases when a feature’s original association with the response variable is nullified. While automatic techniques are available within a logistic regression

Figure 9 gives a high-level overview of how the random forest algorithm works. Briefly, the algorithm takes the full dataset of the letters and their features, and draws random subsamples of letters a given number of times. In the figure, we illustrate with six iterations. In the actual research this was done 2000 times. Each time, about 63% of the letters are included in the random subsample. These are shaded in blue. The remaining 37% of observations that were not included in the subsample make up the test set for the tree, technically called out-of-bag (OOB) observations. These are shaded in green.

A decision tree is then built for each ‘blue’ training set, splitting first on the most important feature that helps to separate response classes. The decision tree algorithm then continually sub-divides the data along linguistic feature lines of successively diminishing discriminative power until a stopping criterion, such as the minimum number of observations in a particular node, is reached. A box provides an intuitive explanation of a decision tree with a worked through hypothetical example.

---

framework for selecting variables e.g. stepwise regression in which variables are sequentially added or removed to see if their addition or removal impacts model fit, such techniques are often unstable, with results affected by the order in which predictors are included or deleted from the model (Strobl et al. 2009).

---

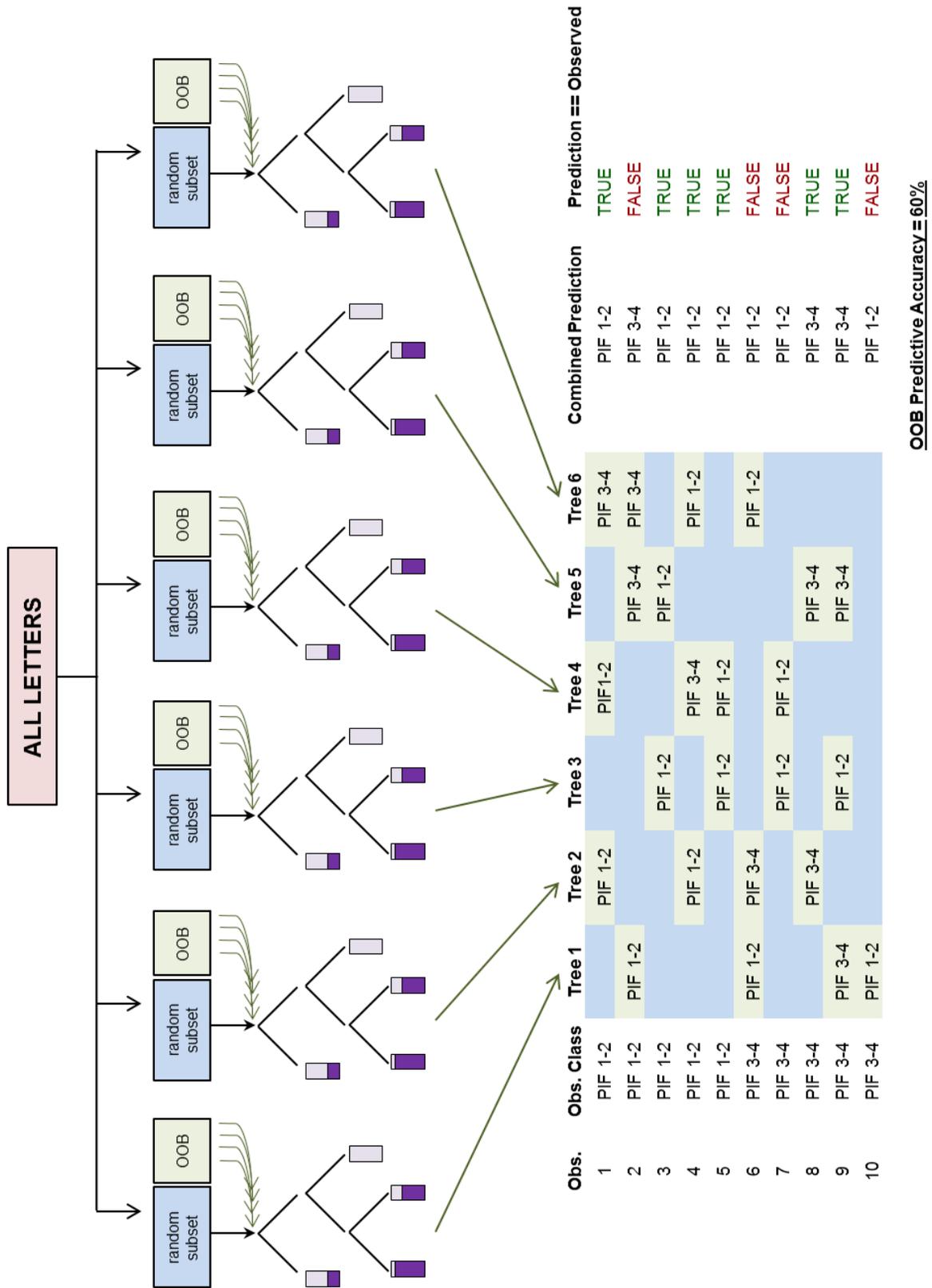
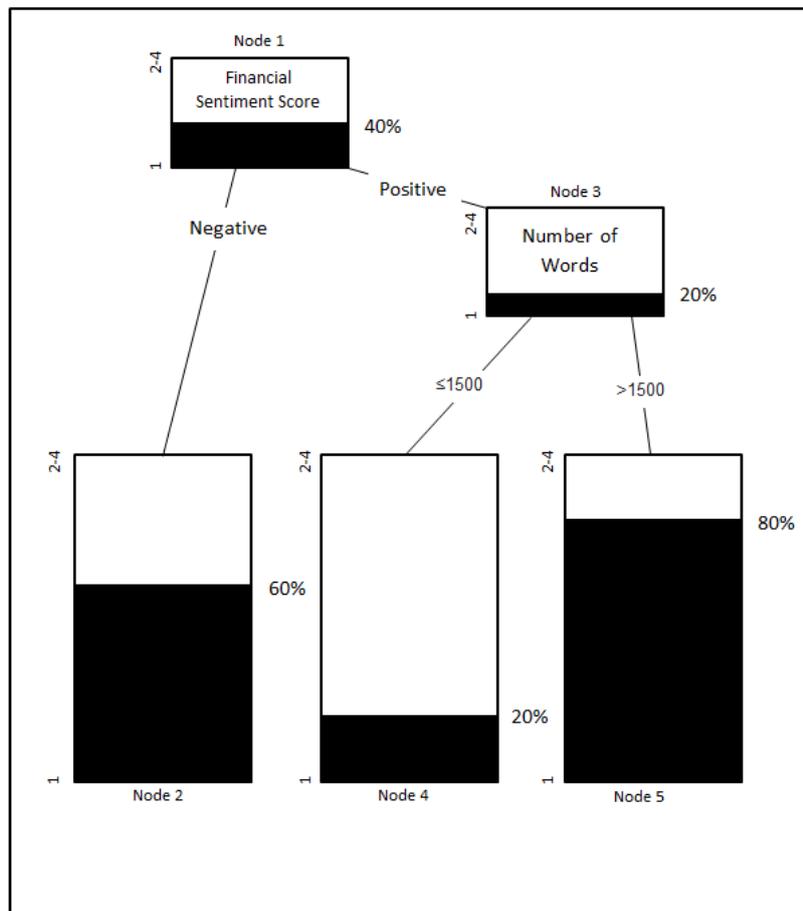


Figure 9: High-level overview of random forest algorithm

### Box: Interpreting decision trees



The plot above is a hypothetical example of a decision tree classifying letters to Category 1 versus Category 2-4 firms. In this tree, the nodes are numbered left-to-right, with the numeration starting at the top node. At each node, a stacked bar plot is produced, showing the proportion of letters (labelled on the right of the bar plot) belonging to the two binned Category classes (labelled on the left of the bar plot).

In this hypothetical example the full dataset is represented at Node 1 with about 40% being Category 1 letters. The algorithm detects the financial sentiment score as the feature with the strongest split-point, and separates the dataset into two parts—one subset containing letters with negative sentiment (Node 2, where 60% are Category 1 letters), and another containing positive sentiment (Node 3, where 20% are Category 1 letters). At Node 2, no further relevant splits are made, so the tree algorithm moves to Node 3 where it splits letters with positive sentiment on the number of words in the letter, separating those letters where the number of words is less than or equal to 1500 (Node 4, where 20% of the letters are Category 1) from those where the number of words is greater than 1500 (Node 5, where 80% of the letters are Category 1). The decision ‘rules’ associated with this tree are thus:

- if the letter’s sentiment is negative, predict ‘Category 1
- if the sentiment is positive and the number of words is greater than 1500, also predict ‘Category 1’
- Otherwise, predict ‘Category 2-4’

To induce more randomness and make the trees more diverse, each time the data is split, only a random selection of the full set of variables is considered to split on. This is usually the square root of the number of features, called *mtry*. In our models, we considered 5 randomly sampled features at each split point i.e. the square root of our full set of 25 features. When all the trees are grown, we have a forest, depicted in the top half of Figure 9.

To assess the performance of the forest (that is, how well it separates the response classes), the OOB observations are then passed through the tree, and the tree assigns a label with its most likely class to each of them. This happens for all the trees in the forest. To exemplify – as shown in the table in Figure 9, in Tree 1, Observation 2 is out-of-bag. It is passed through the tree, and based on its feature values, the tree thinks that it is a PIF1-2 letter. Then, the OOB predictions for a given observation are aggregated, and the winning class is the label that is assigned the most for an observation. So, for example, for Observation 2, the aggregated prediction is PIF1-2 as that is the predicted class in two of the three trees in which it was an out-of-bag observation. Predictive accuracy is then the percent of correct classifications – that is the proportion of times the predicted class matches the observed class.<sup>18</sup>

We assessed the influence of a feature in the model, by using an in-built variable importance procedure. Simplifying somewhat, each feature is given a score based on how much, on average, the predictive accuracy of a single tree in the forest drops when its effect is removed. If out-of-bag accuracy drops a lot, then the feature is useful at discriminating between response classes. If the drop in accuracy is negative/close to zero, this indicates the feature's marginal value add is low or null.

If we run the algorithm different times, we are likely to get slightly different prediction accuracies and (sometimes very) different variable importance rankings, because of the inherent randomness involved in both subsampling the letters and in sampling the features for splitting. In order to ensure that our random forest model and the resulting feature importances are not atypical, we went one step further and built 100 such forests and averaged prediction accuracy. For feature importance, we computed the percent of times a feature showed up as being important across the 100 forest runs. In sum, we simulated 200,000 decision trees (*ntree* = 2,000 times 100 forest runs).

We also used predictions from the forest model to produce dependency plots that allow us to gauge how a linguistic feature relates to, for example, PIF stage, while holding other variables constant.

#### 4. Measuring regulatory proportionality

We now explore the extent to which the PRA's communications to a firm are consistent with the degree of risk that the firm poses to the PRA's statutory objectives, whether that is inherent risk as measured by a firm's category (reflecting its Potential Impact), or imminent risk, as measured by a firm's PIF stage. Our analysis unfolds in three parts. First, we explore how predictable Category and PIF are based on the measures of linguistic complexity, sentiment indicators, directiveness and formality we identified earlier.<sup>19</sup> Second, of these

---

<sup>18</sup> For our PIF and Category models we used a different measure of predictive performance due to data imbalance. See below and, for more detail, Annex 3.

<sup>19</sup> We exclude analysis of forward-lookingness features for two reasons. First, all PSM letters should be forward-looking, regardless of the firm's Category or PIF stage. Second, even when such features are included, they neither enhance the models' performance nor appear in the top-ranking feature sets.

features, we identify the most relevant for discriminating the Category and PIF classes. Third and finally, we identify the direction of the effect.

To facilitate the analysis and to aid interpretation, we grouped observations into two. For Category, we divided the data between Category 1 firms and Category 2–4 firms. This dichotomisation reflects the business reality that Category 1 firms are supervised in a different PRA directorate (Major UK Deposit Takers i.e. MUKDT) from Category 2–4 firms (Banks, Building Societies and Credit Unions i.e. BBSCU). For PIF, we separated firms at PIF stages 1 and 2 from those at PIF stages 3 and 4. This binning is justified on the grounds that firms with a PIF score of 3 and 4 are normally on the PRA’s Watchlist,<sup>20</sup> while those with a PIF score of 1 and 2 are not.<sup>21</sup> Descriptive statistics on all the linguistic features split by PIF and Category are given in Figure 10, Figure 11, Figure 12 and Figure 13.

---

<sup>20</sup> The PRA Watchlist is a list of firms which supervision believes represent a potential risk to the PRA’s statutory objectives. The main purpose of adding a firm to the Watchlist is to escalate their discussion to PRA senior management.

<sup>21</sup> PRA senior management occasionally overrule this when specific firm circumstances are appropriate.

Feature	Category 1 letters			Category 2-4 letters		
	median	mean	sd	median	mean	sd
number of words (in letter)	1919.5	1958.38	482.04	1434	1497.09	536.68
number of section headings (in letter)	10	10.12	4.22	8	8.7	3.65
proportion of acronyms (/number of words) (%)	1.2	1.15	0.27	1.27	1.34	0.63
proportion of numerals (/number of words) (%)	1.18	1.38	0.51	1.74	1.98	0.87
mean sentence length	27.12	26.88	2.71	26.76	26.79	2.5
mean punctuation rate per sentence	2.79	2.87	0.41	2.8	2.83	0.43
mean subordinator rate per sentence	1.31	1.27	0.22	1.38	1.39	0.23
mean verb rate per sentence	4.19	4.23	0.4	4.27	4.29	0.43
financial sentiment score	-0.01	-0.01	0.01	0	-0.01	0.01
proportion of high-risk associated words (%)	0.76	0.76	0.16	0.5	0.55	0.28
proportion of obligatives (/number of words) (%)	0.62	0.63	0.16	1.03	1.05	0.4
proportion of deadlines (/number of words) (%)	0.07	0.07	0.04	0.18	0.2	0.13
proportion of 'please' (/number of words) (%)	0.1	0.1	0.04	0.16	0.19	0.13
ratio of sentence-initial 'please' : sentence-medial 'please'	0.67	0.83	0.53	1.5	1.73	1.04
ratio of sentence-initial 'you' : sentence-medial 'you'	0.26	0.28	0.09	0.28	0.33	0.25
proportion of 1st/2nd personal pronouns (%)	3.44	3.45	1.37	3.64	3.75	1.04
ratio of 'I' : PRA	0.46	0.53	0.3	0.28	0.31	0.21
ratio of 'I' : 'we'	0.07	0.13	0.13	0.06	0.07	0.05
ratio of 'we' : PRA	7.41	6.77	4.24	4.23	5.41	3.97
ratio of 'you' : firm	0.36	0.59	0.61	0.45	0.88	1.16

Figure 10: Descriptive statistics for quantitative features by Category

Qualitative Feature	Category 1 letters		Category 2-4 letters	
	absent	present	absent	present
Presence of appendix	37.50%	62.50%	36.54%	63.46%
Handwritten/typed salutation	handwritten 75%	typed 25%	handwritten 1.92%	typed 98.08%
Generic/named salutation	generic 25%	named 75%	generic 92.31%	named 7.69%

Figure 11: Descriptive statistics for qualitative features by Category

Legend (Linguistic dimension): Complexity; Sentiment; Directiveness; Formality.

Feature	PIF 1-2 letters			PIF 3-4 letters		
	median	mean	sd	median	mean	sd
number of words (in letter)	1439	1478.07	486.94	1819	1801.39	737.38
number of section headings (in letter)	8	8.8	3.67	9	8.83	3.9
proportion of acronyms (/number of words) (%)	1.27	1.35	0.64	1.1	1.21	0.45
proportion of numerals (/number of words) (%)	1.69	1.96	0.89	1.67	1.83	0.72
mean sentence length	26.7	26.67	2.61	27.42	27.43	1.77
mean punctuation rate per sentence	2.75	2.82	0.44	2.92	2.92	0.33
mean subordinator rate per sentence	1.37	1.38	0.24	1.45	1.39	0.19
mean verb rate per sentence	4.25	4.27	0.44	4.29	4.36	0.35
financial sentiment score	0	-0.01	0.01	-0.01	-0.01	0.01
proportion of high-risk associated words (%)	0.48	0.51	0.24	0.86	0.83	0.32
proportion of obligatives (/number of words) (%)	1.04	1.08	0.37	0.64	0.71	0.41
proportion of deadlines (/number of words) (%)	0.17	0.2	0.13	0.12	0.14	0.1
proportion of 'please' (/number of words) (%)	0.16	0.2	0.13	0.12	0.13	0.07
ratio of sentence-initial 'please' : sentence-medial 'please'	1.5	1.7	1.08	1.25	1.51	0.83
ratio of sentence-initial 'you' : sentence-medial 'you'	0.27	0.33	0.25	0.28	0.3	0.17
proportion of 1st/2nd personal pronouns (%)	3.63	3.76	1.06	3.73	3.54	1.12
ratio of 'I' : PRA	0.29	0.33	0.23	0.3	0.34	0.19
ratio of 'I' : 'we'	0.06	0.07	0.04	0.07	0.11	0.12
ratio of 'we' : PRA	4.15	5.58	4.08	4.92	5.13	3.52
ratio of 'you' : firm	0.43	0.88	1.2	0.72	0.77	0.62

Figure 12: Descriptive statistics for quantitative features by PIF

Qualitative Feature	PIF 1-2		PIF 3-4	
	absent	present	absent	present
Presence of appendix	32.98%	67.02%	55.56%	44.44%
Handwritten/typed salutation	handwritten 2%	typed 98%	handwritten 33.33%	typed 66.67%
Generic/named salutation	generic 91%	named 9%	generic 66.67%	named 33.33%

Figure 13: Descriptive statistics for qualitative features by PIF

Legend (Linguistic dimension): Complexity; Sentiment; Directiveness; Formality.

## Differences between PSM letters sent to different categories of firms

Overall, our random forest model for Category has a mean out-of-bag<sup>22</sup> predictive accuracy (C-statistic) of 0.9, a clear improvement on the no-information ‘guess rate’ of 50%. We explain the details of and our motivations for using this statistic in the annex. Suffice it to say here, C is the probability that a randomly chosen Category 1 firm letter will be assigned a higher predicted probability of being a Category 1 firm letter compared with a randomly chosen Category 2-4 letter. Readers familiar with standard statistical methods can think of this as roughly an R squared measure. In the machine learning literature a C-statistic this high is considered “outstanding” (Hosmer et al. 2013: 177). Given that predictive accuracy evaluated on the OOB data is merely an estimate of out-of-sample performance, we also assessed how well the model predicts unseen data. Testing the algorithm on the 2016 vintage of PSM letters, we found predictive accuracy to be perfect ( $C = 1$ ).<sup>23</sup> In short, our random forest shows that Category 1 firms are on the whole linguistically different from those written to Category 2-4 firms.

Figure 14 shows that seven of the 23 linguistic features were identified as salient in all 100 runs of the random forest. These linguistic features are shaded orange. Strobl and her co-authors (2009) suggest that variables with importance scores that are negative, zero, or barely positive are uninformative and can be ignored.<sup>24</sup> If the variable’s value is significantly positive, the variable is considered potentially informative. One can think of values above this threshold as being “statistically significant”, in an extremely loose sense of the term. In terms of implementation, for each of the 100 random forests grown, we computed variable importance rankings for each feature, and tagged each feature as being above or below the threshold suggested by Strobl and co-authors. We plot the variable importance of the seven linguistic features that are above this threshold in Figure 15. Figure 16 plots the seven key linguistic features identified and the direction of the effect. These dependency plots show how the model’s predicted probability of a letter being a Category 1 letter changes with increasing values of that specific linguistic feature, holding all other covariates at their median values (for quantitative features) or mode values (for qualitative features). For presentational purposes, we give predicted probability on the y-axis in percentage terms.<sup>25, 26</sup>

---

<sup>22</sup> Out-of-sample performance estimates based upon the OOB data have been shown to be roughly equivalent to those based on cross-validation (Hastie et al. 2009: 593). For the sake of completeness, however, we also performed in-domain ‘leave-one-out cross-validation’ in which we build  $n$  models on  $n - 1$  observations and predict the held-out observation on each iteration. For this classifier,  $C_{(CV)}$  is 0.91, which is a slightly better estimate of the out-of-sample accuracy than the OOB.

<sup>23</sup> It may seem strange that predictive accuracy of the model based on completely unseen data should be higher than that estimated using the OOB data or that based on cross-validation. However, it should be noted that the test dataset is less variable than the training dataset on which the OOB accuracy score was computed, especially with respect to those features that drive Category classification. In other words, the unseen data is easier to classify than the OOB data.

<sup>24</sup> Formally, we state this threshold as  $t \leq |\min(V)|$  where  $t$  is the threshold below which a predictor is uninformative, and  $V$  is the set of all 25 importance scores for a given forest iteration.

<sup>25</sup> Because the predicted probability of a letter being a Category 1 letter is vanishingly small when all variables are held at this level ( $\approx 0.15\%$ ), and because the linguistic variables do not typically show dramatic effects, we limit the range of the y-axis to  $[0\% - 10\%]$  to allow the effect differences to show up more clearly.

<sup>26</sup> Note also that the differences seen in these plots are so tiny (especially apparent in Panel G), that it is unlikely that these features’ main effects are determinate on their own. Instead, it is likely that these variables participate in complex interactions with each other and other features. These cannot be revealed in these two-dimensional dependency plots because the values of the other interacting variables are held constant in producing the probability estimates.

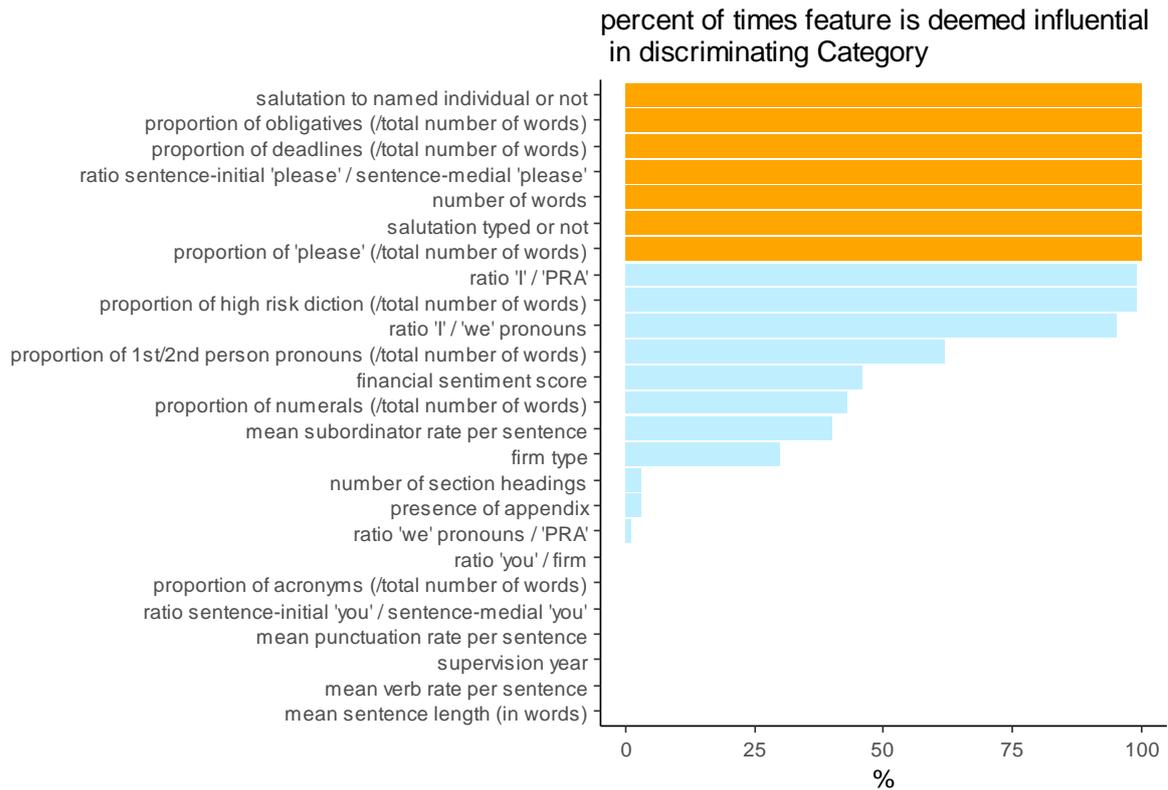


Figure 14: Percent of times a feature is deemed influential at discriminating Category in the 100 forest runs, ordered vertically from most to least influential. Features that are detected as being influential in all runs are shaded in orange. Other features are shaded blue. If no bars appear, this indicates that these features never appeared as influential in any of the forest runs.

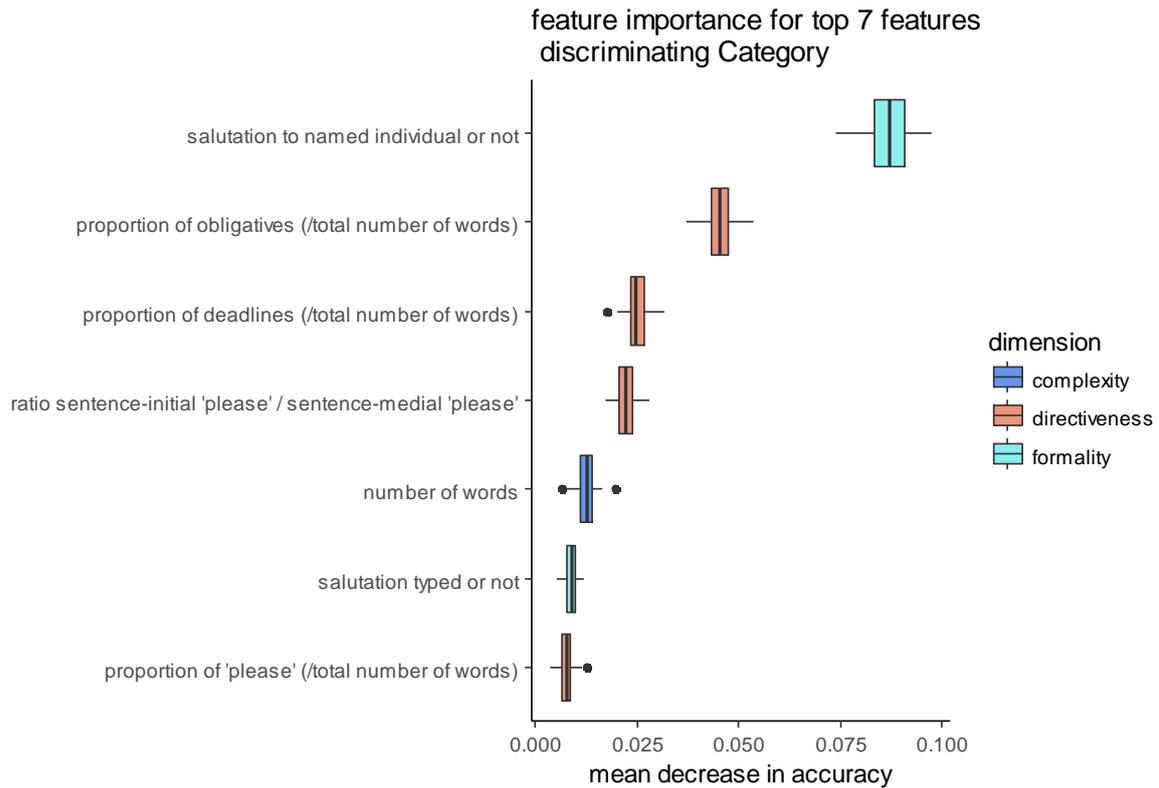


Figure 15: Plot showing the summary of variable importance for the 7 top-ranking predictors of Category over 100 forest iterations, colour-coded by type of linguistic feature. Please see Section 3 for definitions.<sup>27</sup>

<sup>27</sup> These boxplots are flipped so that the feature's name is easier to read, with the numerical information consequently appearing on the x-axis. For each feature, the x-axis denotes the average amount by which the accuracy of a single tree in the forest drops when the feature of interest is disassociated with the response variable. For instance, when the proportion of obligatives is disassociated with Category, then an individual tree's C-statistic drops by 0.05 on average, thus indicating that this feature has some discriminatory power. The greater the decrease in accuracy for a particular feature, the more important the variable is and the further to the right it appears in the plot.

Dependency plots for Category

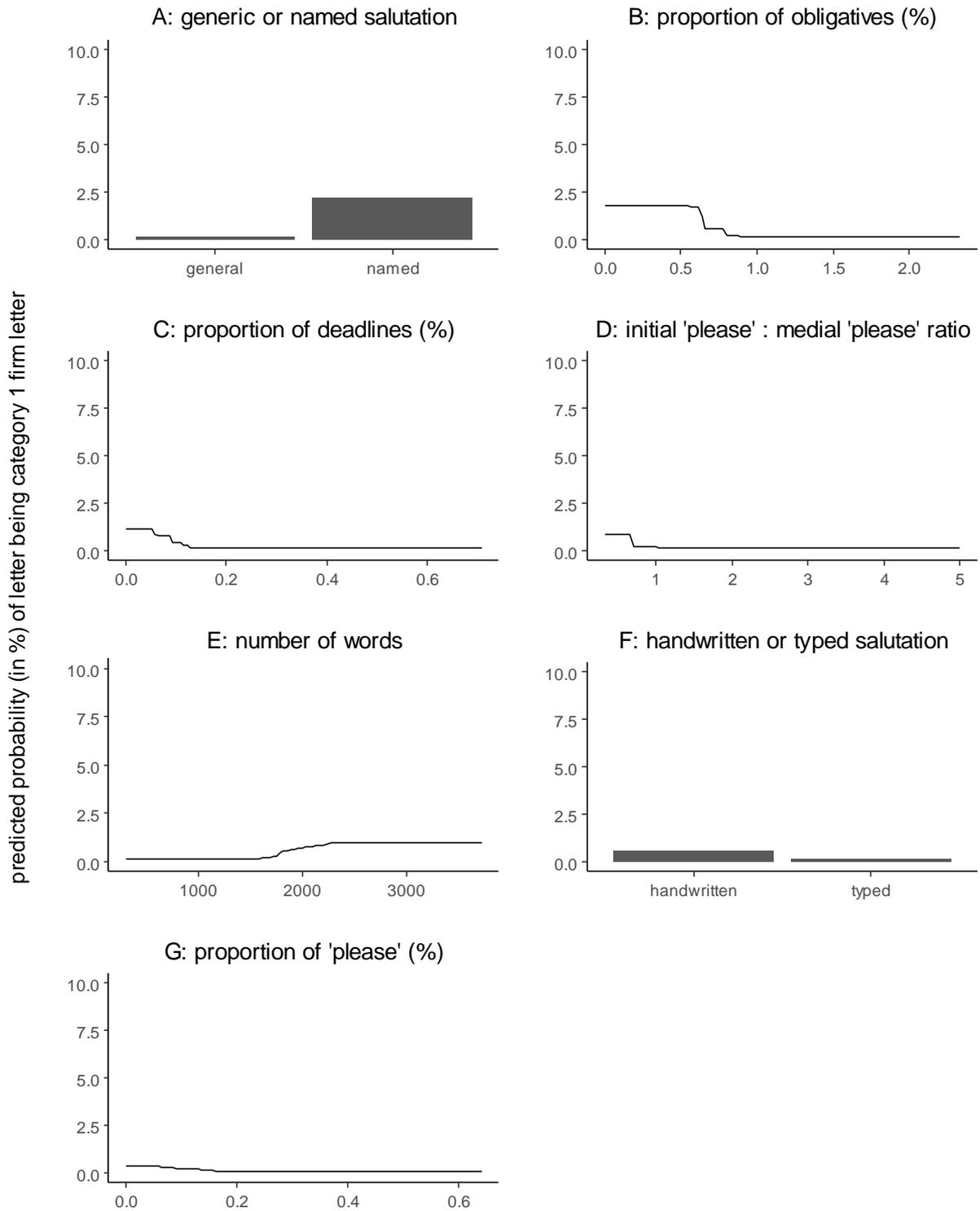


Figure 16: Target variable dependency plots on each of the 7 top-ranking features for Category classification. The black line (or bars) gives the predicted probability (in %) of a letter being a Category 1 firm letter at varying values of the linguistic feature of interest.

Taking these three plots together, we observe that the most important set of linguistic features relate to directiveness. These include the proportion of obligatives (out of the total number of words), the proportion of deadlines (out of the total number of words), the ratio between ‘please’ occurring in sentence-initial position versus its occurrence in a sentence-medial position,<sup>28</sup> and the proportion of ‘please’ overall (out of the total number of words). To spell out their effects, we see in Panel C in Figure 16 that as the mentions of deadlines increase, the predicted probability of a letter being Category 1 decreases. Similarly in Panel D in the same figure, as the ratio of initial ‘please’ to medial ‘please’ in a letter increases, the predicted probability of a letter being Category 1 decreases. The same pattern is weakly discernible in Panel G. Collectively, these plots suggest that Category 1 letters are identifiable by less directiveness compared with Category 2–4 letters. In a nutshell, we found the PSM letters to Category 2-4 firms to be more directive than those written to Category 1 firms.

Two linguistic measures of formality are relevant — whether the letter’s salutation is addressed to a named individual, and whether the salutation is typed. Panel A in Figure 16 shows that named salutations (e.g. “*Dear John*”) make it more probable that the letter is a Category 1 firm letter, and Panel F in the same figure shows that handwritten salutations also increase prediction for a Category 1 firm letter.

Only one linguistic feature pertaining to complexity was found to be influential – the number of words. As Panel E in Figure 16 demonstrates, letters are longer to Category 1 firms than the letters sent to Category 2-4 firms.

It is notable that none of our sentiment based features turned out to be relevant. In general, we found that the sentiment of PSM letters was neither ‘positive’ nor ‘negative’, as can be seen in Figure 17 and Figure 18. In other words, their tone was neutral, as one would expect in a professional ‘business-to-business’ correspondence.

---

<sup>28</sup> We include this linguistic feature because sentence-initial ‘please’ is more direct (blunt) than when it occurs in other sentential positions (Danescu-Niculescu-Mizil et al. 2013). See the annex for an example.

### Sentiment with financial context - Year 2014

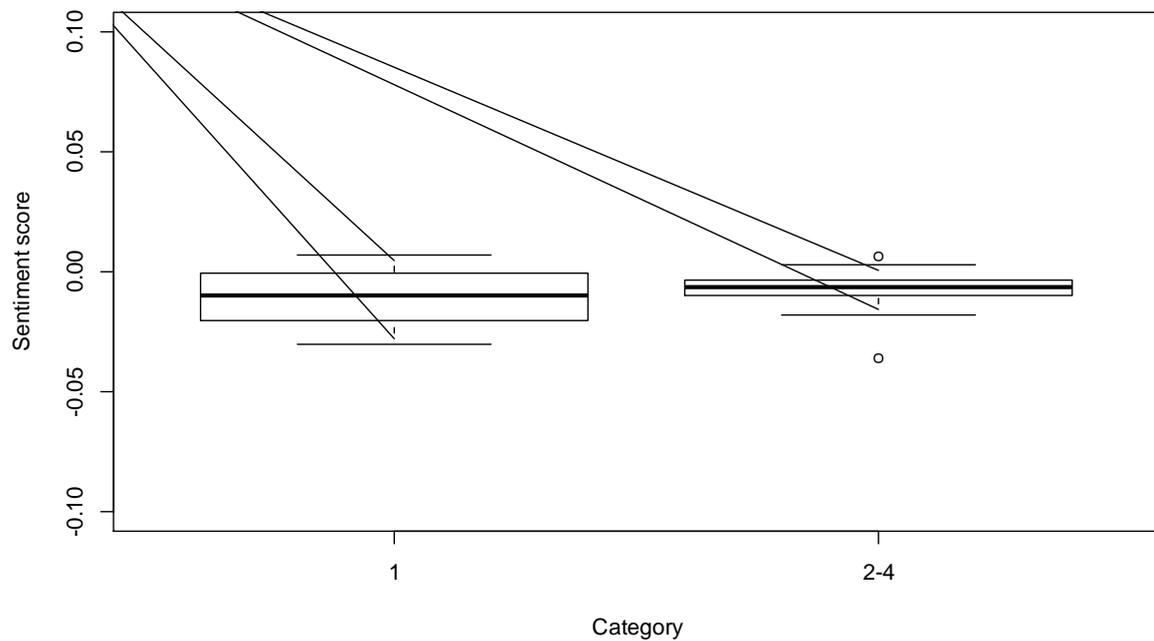


Figure 17: Sentiment scores for all firms in 2014. For details on methodology please see Annex 2. The sentiment scores are the percentage difference between positive and negative words. Even though most scores are slightly negative for the year 2014, the balance is very close to zero. This can be interpreted as neutral sentiment.

### Sentiment with financial context - Year 2015

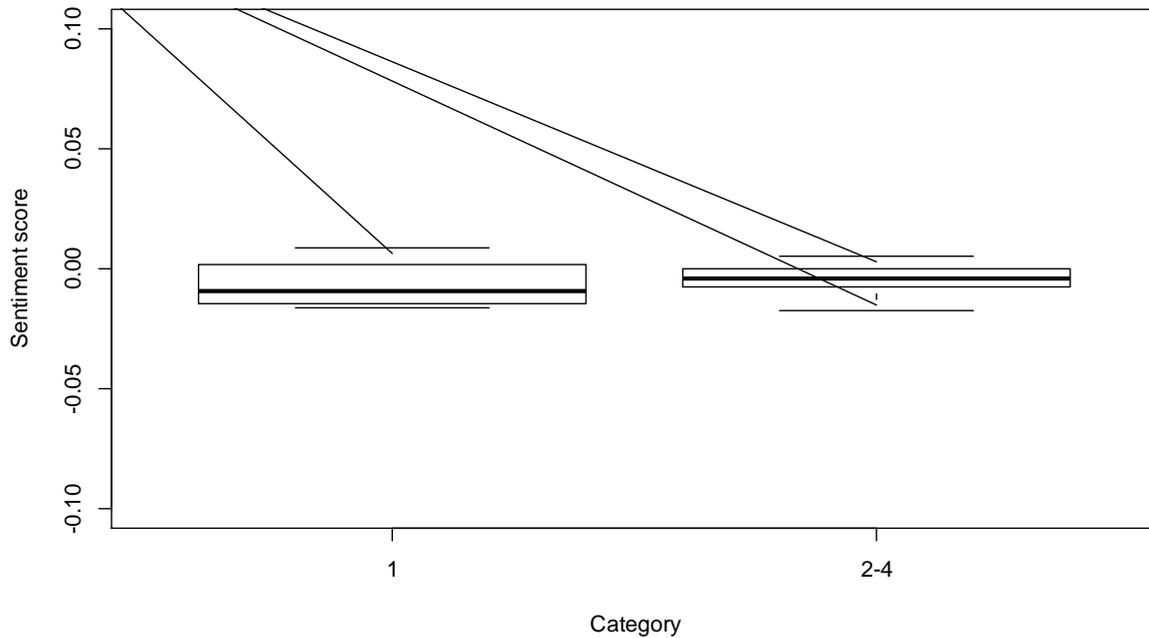


Figure 18: Sentiment scores for all firms in 2015.

We conclude our analysis of Category by including a table of relevant features, their higher-level grouping, and the direction of the effect for Category 1 firm letters versus Category 2-4 firm letters (Figure 19).

Complexity	length of letter (in words)	longer
Directiveness	proportion of obligative words in letter (out of total number of words)	fewer
	proportion of deadlines in a letter (out of total number of words)	fewer
	proportion of 'please' in a letter (out of total number of words)	fewer
	ratio of sentence-initial 'please' count to sentence-medial 'please' count	fewer
Formality	whether the salutation is handwritten or not	handwritten
	whether the salutation is to a named individual or not	named

Figure 19: Influential features in the analysis of Category and effect directions for Category 1 firms (compared with Category 2-4 firms)

## Differences between PSM letters to firms at different PIF stages

We now compare the writing styles of letters sent to firms at different PIF stages. Recall that PIF refers to a firm's proximity to resolution, with firms at the PIF 3-4 stage closer to resolution. Like our random forest for Category, our model fit for PIF was excellent, with a mean OOB C-statistic of 0.84.<sup>29</sup> We also obtained outstanding discrimination on our 2016 test dataset, with a mean C-statistic of 0.87, validating the generalizability of the model to new data. Our model found linguistic features that differentiate PSM letters sent to PIF 1-2 firms from those sent to PIF 3-4 firms. Figure 20 plots the percentage of time a feature is above the required significance threshold across all 100 forest iterations. We find four linguistic features are useful in discriminating between the two PIF classes in all 100 runs. In addition, the control variable as to whether a firm is a bank or a building society is significant. This reflects the fact that banks in this sample are more likely to belong to a particular PIF class compared with building societies. Figure 21 plots the relative contributions of these five features.

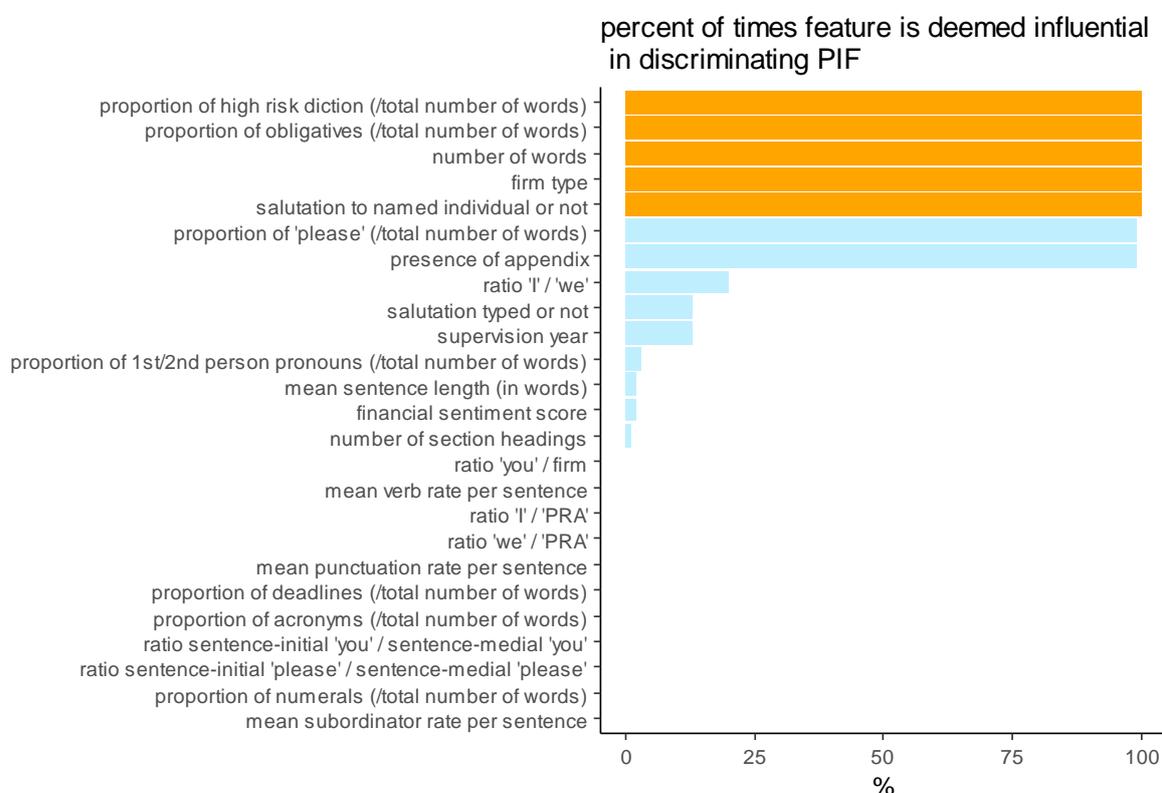


Figure 20: Percent of times a feature is deemed influential at discriminating PIF stages in the 100 forest runs, ordered vertically from most to least influential. Features that are detected as being influential in all runs are shaded in orange. Other features are shaded blue. If no bars appear this indicates that these features never appeared as influential in any of the forest runs.

<sup>29</sup> As a further measure of robustness, we evaluated performance through in-domain leave-out-one cross-validation. The cross-validated value for  $C_{(CV)}$  is 0.84, identical to the OOB estimate.

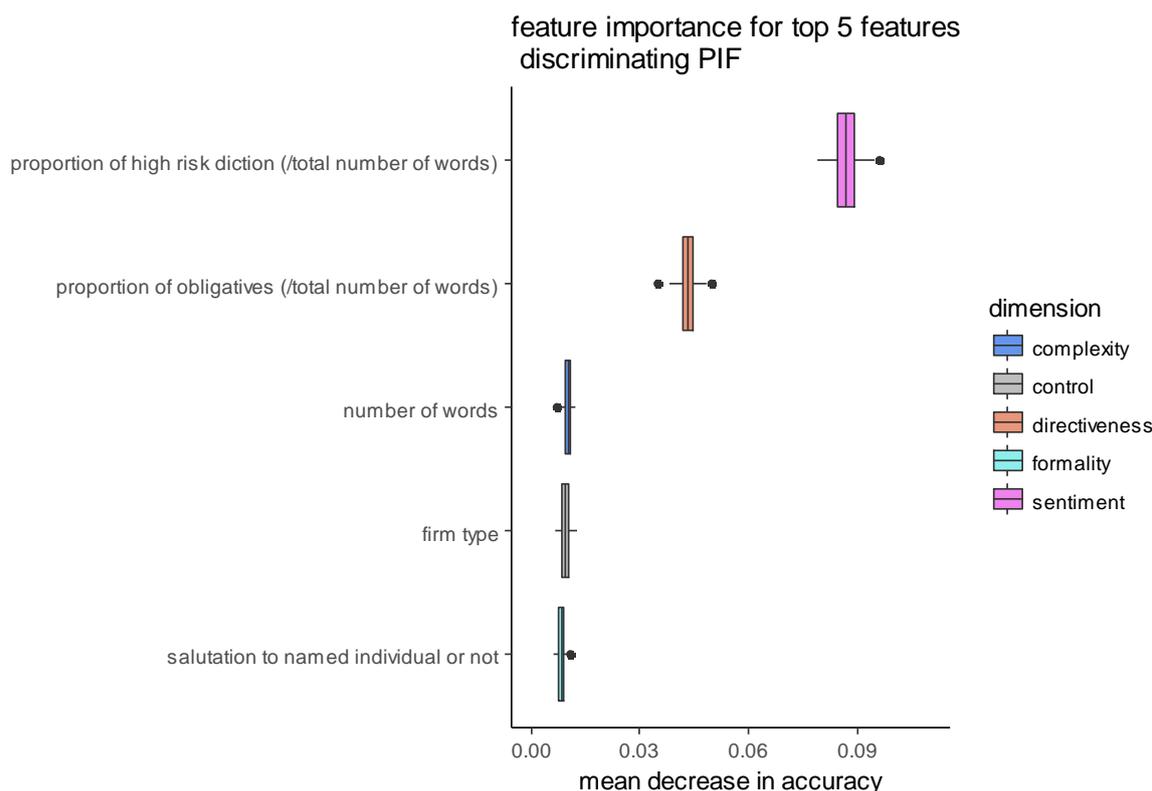


Figure 21: Plot showing the summary of variable importance for the 5 top-ranking predictors of PIF over 100 forest iterations, colour-coded by the dimension of proportionality. The x-axis shows the decrease in accuracy (as given by the C-statistic) that results from disassociating the feature with the response.

Figure 22 contains our dependency plots. Panel A shows that the predicted probability of a PIF 3-4 letter increases sharply after the percent of high risk diction in a letter passes beyond 0.75%. Unsurprisingly, letters to firms that are higher risk contain a greater normalised frequency of high risk diction. But counterintuitively, Panel B shows that, as the proportion of obligatives in a letter increases, the predicted probability of a PIF 3-4 letter decreases. One might have suspected that letters to higher risk firms would be relatively more directive. However, the issues that higher risk firms face are much more complex and interwoven compared with those faced by lower risk firms. Complex issues often cannot be readily addressed through a simple prescriptive imperative in a summary document, and it is likely that other communications are sent to the firm in addition to the PSM letter addressing the issues in a more nuanced way.<sup>30</sup>

<sup>30</sup> We caution that all the features in the dependency plots are very weak signals on their own, instead pointing to the strong likelihood of complex interactions between them.

Dependency plots for PIF

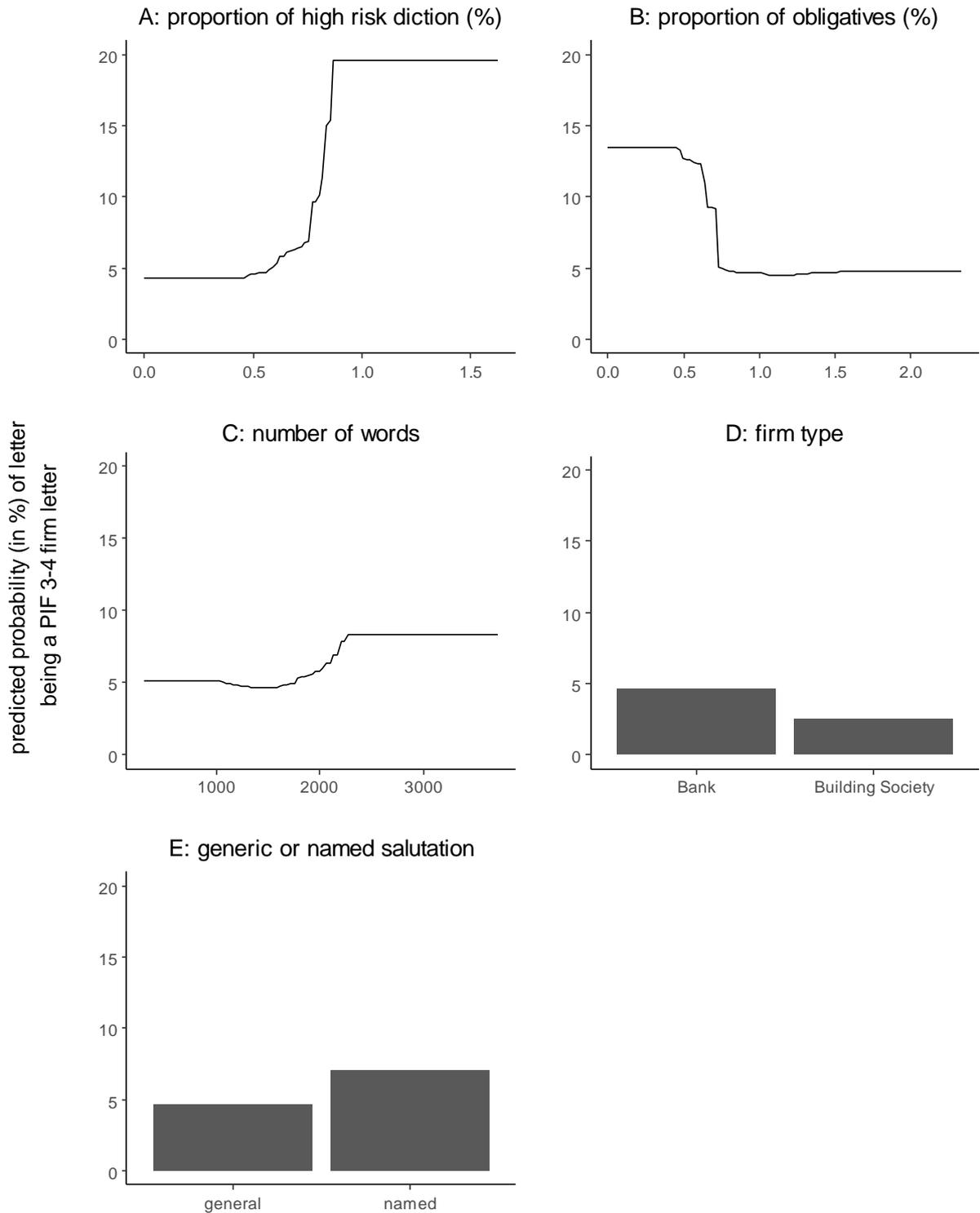


Figure 22: Target variable dependency plots on each of the 5 top-ranking features for PIF classification. The black line (or bars) gives the predicted probability (in %) of a letter being a PIF 3-4 firm letter at varying values of the linguistic feature of interest.

## 5. How supervisory communication has changed— PRA versus FSA letters

One way to understand the PRA’s current supervisory approach is that it is, at least in part, a response to some of the criticisms levelled at the FSA following the financial crisis. For instance, we earlier noted that the new approach emphasises judgement. Whether intended or not, this implicitly draws a contrast between it and what some perceived as ‘box ticking’ at the FSA. Similarly, the emphasis on being forward-looking reflects concern with low probability but high impact events that the previous supervisory regime failed to foretell. One way to understand the extent to which there has been a shift in supervisory regime 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.

### Random forest model

We compared the two batches of letters along 24 of the linguistic features we posited earlier, with the exception of section headings, which we analyse separately below. Figure 23 contains descriptive statistics on the median, mean, and standard deviations of the quantitative linguistic features. Figure 24 presents these statistics.

Quantitative Feature	PSM			ARROW		
	median	mean	sd	median	mean	sd
number of words (in letter)	1490.5	1515.05	472.99	1062	1231.44	595.08
proportion of acronyms (/number of words) (%)	1.35	1.38	0.6	0.79	0.86	0.39
proportion of numerals (/number of words) (%)	1.85	2.05	0.94	1.23	1.27	0.41
mean sentence length	26.99	26.88	2.41	24.09	24.51	2.45
mean punctuation rate per sentence	2.82	2.82	0.39	2.21	2.28	0.46
mean subordinator rate per sentence	1.37	1.4	0.25	1.22	1.23	0.22
mean verb rate per sentence	4.34	4.33	0.43	3.82	3.84	0.47
financial sentiment score	0	0	0.01	-0.01	-0.01	0.01
proportion of high-risk associated words (%)	0.57	0.58	0.26	0.45	0.49	0.27
proportion of obligatives (/number of words) (%)	1	0.99	0.33	0.7	0.71	0.25
proportion of deadlines (/number of words) (%)	0.19	0.2	0.13	0.1	0.12	0.07
proportion of 'please' (/number of words) (%)	0.16	0.2	0.14	0.29	0.31	0.14
ratio of sentence-initial 'please' : sentence-medial 'please'	1.75	1.76	1.2	1	1.68	1.27
ratio of sentence-initial 'you' : sentence-medial 'you'	0.25	0.31	0.24	0.27	0.28	0.09
proportion of 1st/2nd personal pronouns (%)	3.44	3.58	1	4.55	4.89	1.6
ratio of 'I' : PRA (or FSA, as appropriate)	0.2	0.26	0.19	0.33	0.47	0.34
ratio of 'I' : 'we'	0.06	0.08	0.08	0.07	0.07	0.04
ratio of 'we' : PRA (or FSA, as appropriate)	3	3.93	2.9	5.88	7.91	6.93
ratio of 'you' : firm	0.4	0.78	0.97	0.82	1.17	0.92
proportion of future-oriented sentences (%)	32.09	32.47	7.48	22.5	23.04	6.94
proportion of non-past tensed verbs (/tense-marked verbs) (%)	78.12	77.8	7.14	78.57	78.01	6.3

Figure 23: Summary statistics for the quantitative features by letter type (PSM vs. ARROW). Legend (Linguistic dimension): Complexity; Sentiment; Directiveness; Formality; Forward-lookingness

Qualitative Feature	PSM		ARROW	
	absent	present	absent	present
Presence of appendix	33.93%	66.07%	0%	100%
Handwritten/typed salutation	8.93%	91.07%	0%	100%
Generic/named salutation	87.50%	12.50%	98.18%	1.82%

Figure 24: Summary proportions for the qualitative features in PSM vs. ARROW. Legend (Linguistic dimension): Complexity; Formality.

As before, we conducted a test for differences using random forests. The mean out-of-bag predictive accuracy rate was 91%.<sup>31</sup> Thus we find that pre- and post-crisis supervisory correspondence styles are measurably different. Figure 25 plots the variable importance of the linguistic features we posited and measured. We see that 19 features are detected as being potentially informative in all 100 variable importance iterations. For these 19 strongest performing predictors, we plot their relative strengths in Figure 26.

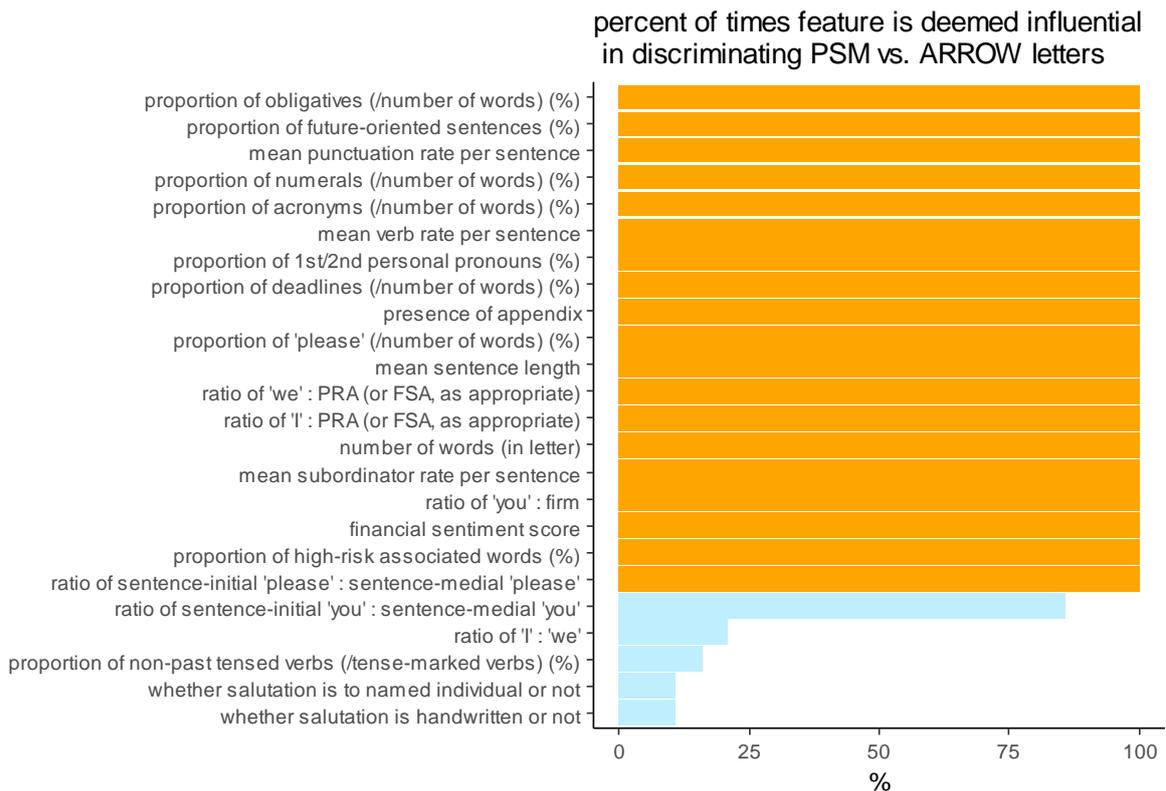


Figure 25: Percent of times a feature is deemed influential at discriminating ARROW from PSM letters in the 100 forest runs, ordered vertically from most to least influential. Features that are detected as being influential in all runs are shaded in orange.

<sup>31</sup> Here we use the percentage of correctly classified out-of-bag observations as our performance metric instead of C because the dataset is roughly balanced with approximately an equal number of observations of the two types of letters. As a further robustness check, we performed leave-out-one cross-validation, according to which the out-of-sample accuracy was estimated at 91%. This value, it will be noted, is the same as the OOB estimate.

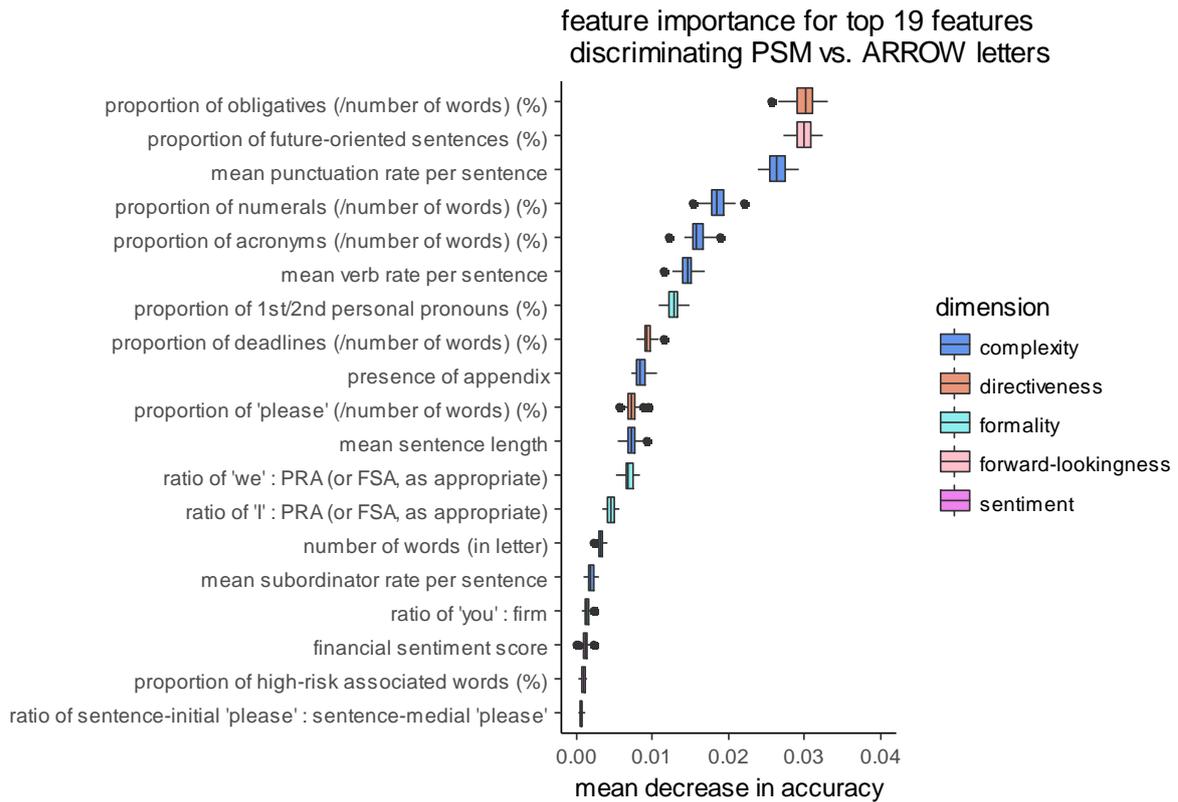


Figure 26: Plot showing the summary of variable importance for the top-ranking predictors of letter period over 100 forest iterations. The x-axis indicates the decrease in accuracy resulting from disassociating a particular feature with the response.

We find that the most important variables are the proportion of obligative structures and the proportion of future-oriented sentences. Although much weaker, the financial sentiment score, the proportion of high-risk associated words, and the number of times ‘please’ occurs sentence-initially as opposed to sentence-medially, still contribute to predictive accuracy.

Figure 27 displays dependency plots for the 19 linguistic features identified as salient for distinguishing between PSM and ARROW letters. As noted earlier, these show the model’s predicted probability of a letter being classified as a PSM letter for the full range of values of a given feature, while holding the values of the other features at their median values (for quantitative variables) or mode level (for qualitative variables). Again these plots allow us to gauge how much the feature contributes to predictive accuracy, or, interpreted in other terms, the way in which a linguistic feature relates to letter type.

Dependency plots for PSM vs. ARROW

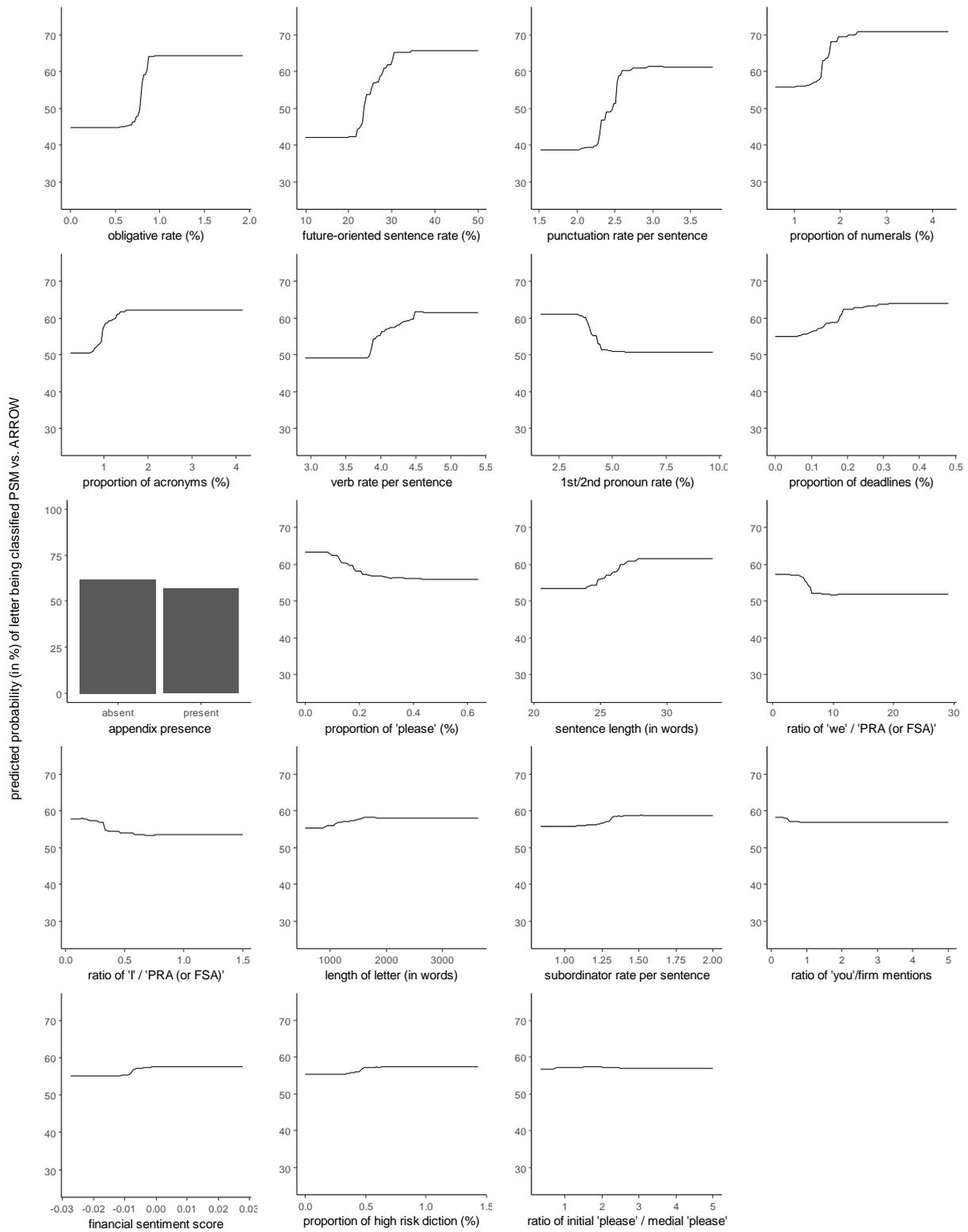


Figure 27: Target variable dependency plots on each of the 19 top-ranking features for ARROW vs. PSM classification. The y-axis gives the predicted probability (in %) of a letter being a PSM letter at varying values of the linguistic feature of interest.

The top-left plot in Figure 27 shows that as the proportion of obligatives increase in a letter, the predicted probability of a letter being a PSM letter increases. The other plots can be read similarly. They show that in the PSM letters, there is a relatively higher rate of future-oriented sentences and deadlines, as well as punctuation marks per sentence, numerals, acronyms, verbs per sentence, and longer sentences compared to FSA ARROW letters. By contrast, there is a relatively lower rate of ‘please’ and local personal pronouns such as ‘I’ and ‘you’. Note that dependency on the presence of an appendix and the last eight features in Figure 27 is hard to discern — the lines are relatively flat. As pointed out by Hastie and co-authors (2009: 373), this indicates that these variables do not show strong main effects in themselves but may be taking part in (higher-order) interactions with the other top-ranking features. These interactions are not easily displayed in two-dimensional dependency plots.

Looking at the results in the round, we detect a stylistic shift between ARROW letters and PSM letters, arguably reflecting a change in supervisory approach. Future-oriented diction is a key linguistic discriminator between the two letter types. One of the critiques of the FSA was that it responded to risks only after they had crystallised. By contrast, the PRA aspires to be more pro-active. Based on our analysis, it appears to be so. We also find that PSM letters are more complex than the FSA’s ARROW letters at the sentence level, in terms of punctuation rate, verb rate, sentence length, and subordinator rate. They are also more complex at the document level, in terms of acronym proportion, numeral proportion, and letter length. The fact that PRA letters are longer perhaps reflects its judgment-based approach and the need to explain its rationale for those judgments. To this point, the greater rate of subordinate clauses perhaps indicates that more thorough explanations are being given in the letters as to the rationale behind regulatory action. In addition, the greater frequency of acronyms and numeric information has increased letter complexity. The higher rate of obligatives and deadlines in PSM letters may indicate that the PRA is, on the whole, being more directive in its communication to firms than was the FSA. Finally, the PRA seems to be adopting a more detached stance in its narrative, preferring more formal styles of reference, such as a reduced rate of 1st (‘I’, ‘we’) and 2nd person pronouns (‘you’).

## Differences in discursive content

To complement our random forest model of linguistic features, we examined the content of the two letter types, using the section headings as proxies for their topics.<sup>32</sup> The section headings were compiled from all the ARROW letters and the most recent vintage of PSM letters in our core sample (2015)<sup>33</sup>. We then grouped these section headings into larger, overarching, standardised categories. For example, section headings on the ‘Net Stable Funding Ratio’ and ‘Liquidity Coverage Ratio’ would be slotted together under the meta-heading ‘Liquidity.’<sup>34</sup> We also left aside generic section headings such as ‘Overall assessment’ and ‘Confidentiality and response to this letter’ from the final list analysed. It is notable that ARROW letters had a much higher percentage of generic section headings compared to PSM letters (Figure 28). We think this means that the PRA is taking a more tailored approach to communication with firms than did the FSA.

---

<sup>32</sup> We tried topic modelling as well but found the results uninformative.

<sup>33</sup> For the sake of completeness, we also produced section heading charts for the first vintage of PSM letters in our core sample (2014) and for the 2016 test data. These are given in Annex 4.

<sup>34</sup> Both the liquidity coverage ratio and net stable funding ratio are measures introduced by Basel III to measure liquidity risk.

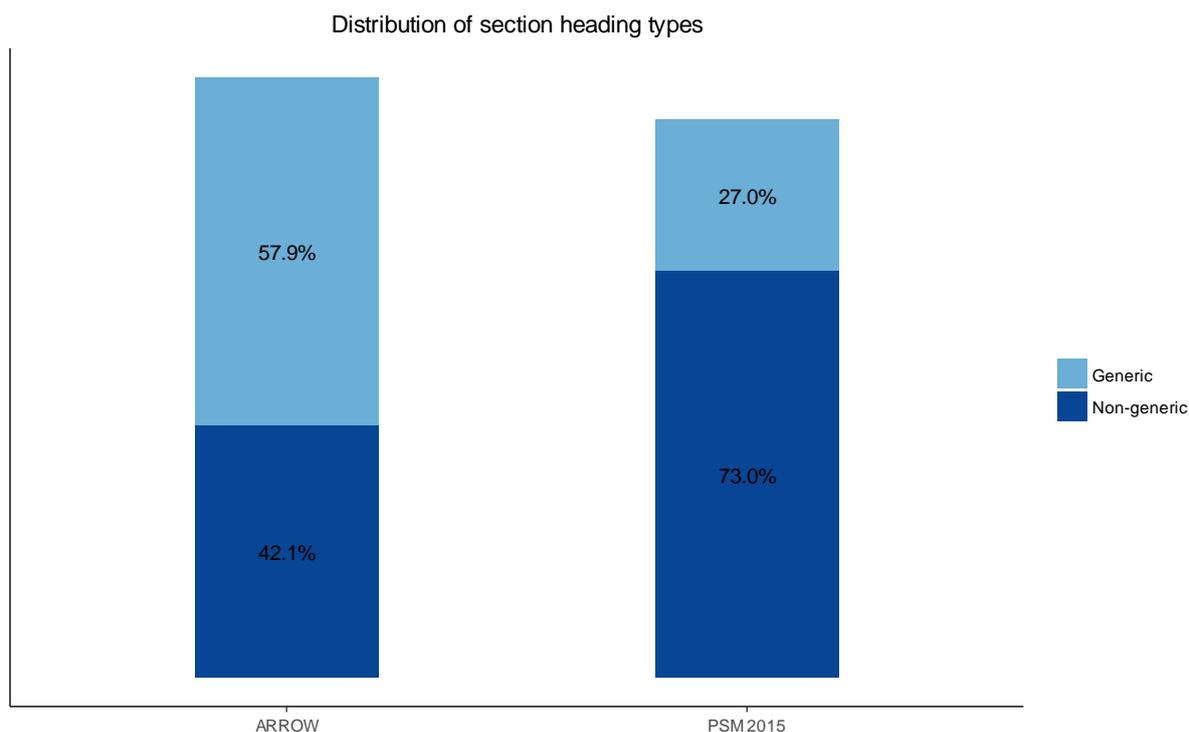


Figure 28: Distribution of section heading types for ARROW and 2015 PSM letters

The standardised, *non-generic* section headings were then subjected to frequency analysis. Figure 29 and Figure 30 present these results in graphical form. Note that we have classified some headings as ‘other non-generic headings’ to protect confidentiality, in cases where these headings appeared in just a few letters.

Among the top ten recurring section headings, ‘Liquidity’ and ‘Recovery and Resolution Planning’ featured prominently in PSM letters but were absent in ARROW letters. This reflects the prominence of these topics on the post-crisis supervisory agenda, and their relative neglect pre-crisis. Equally, ‘Compliance’ and ‘Treating Customers Fairly’ do not appear as section headings in PSM letters but appear often in ARROW letters. This reflects the fact that conduct regulation is now the responsibility of the FCA, not the PRA, whilst the FSA’s statutory objectives included consumer protection and reducing financial crime.<sup>35</sup>

<sup>35</sup> The FSA’s other statutory objectives were to maintain market confidence and financial stability.

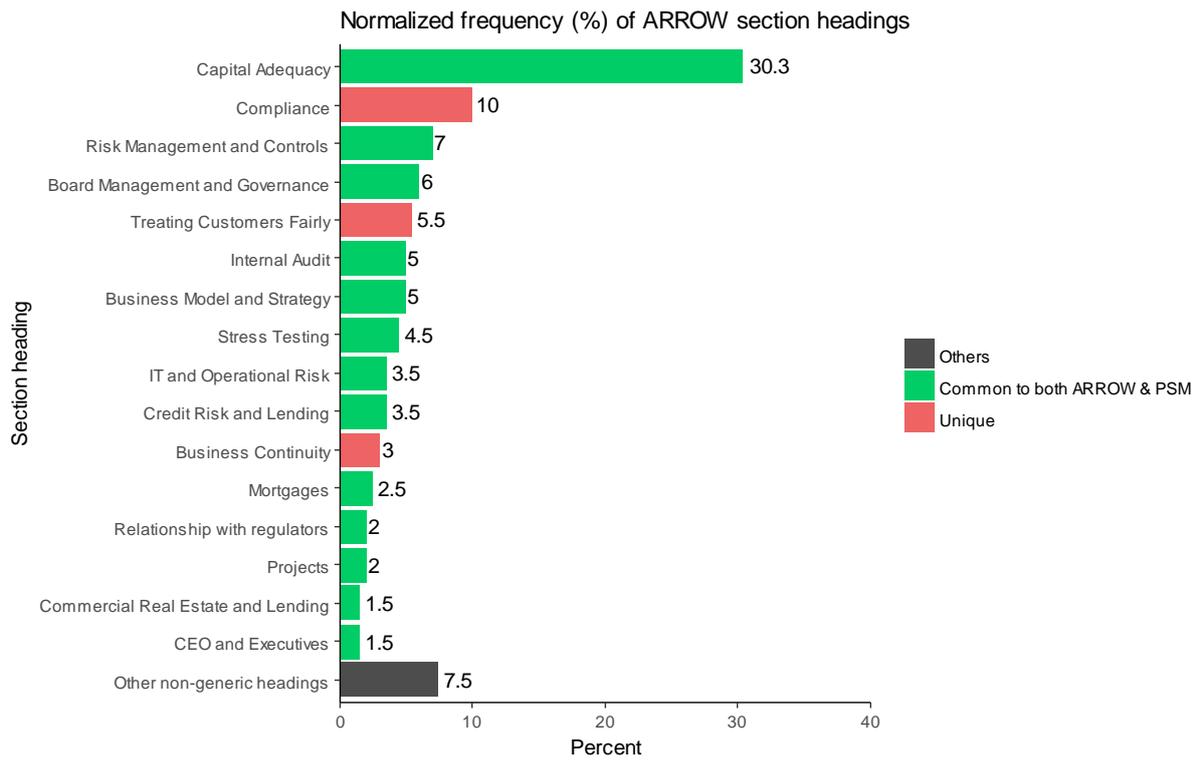


Figure 29: Normalized frequency bar plot of ARROW section headings

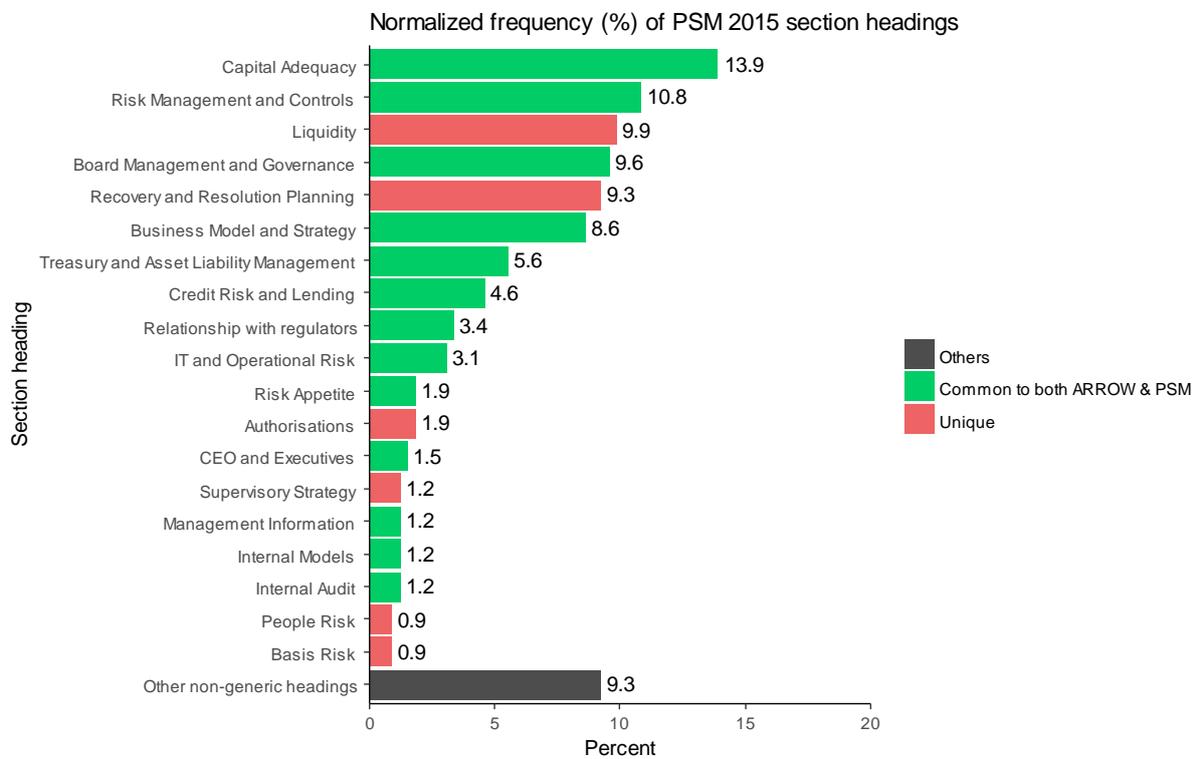


Figure 30: Normalized frequency bar plot of PSM 2015 section headings

We also analysed the proportion of each section heading relative to all section headings in the PSM and ARROW letters. We gathered the section headings common to both PSM and ARROW letters and tracked changes in the proportion of each. Figure 31 shows that Capital Adequacy as a section heading has fallen most significantly in relative proportion by around 16% from the ARROW to PSM letters. We think this decline relates to the fact that PSM letters cover a wider terrain of risks, and that supervisory messages about capital now often come in a separate communication related to stress testing and supervisory review and evaluation process (SREP) of firms individual capital adequacy assessment plans (ICAAPs). For the largest firms with affiliates in other EU countries, such communication is also subject to a Joint Risk Assessment and Decision (JRAD) process involving relevant EU and national competent authorities. It may also be the case that capital is discussed within other sections besides a literally named capital adequacy section.

While there has been a *relative* decline in the discussion of capital in PSM letters, there has been an increase in the proportion of sections related to Board Management and Governance, Risk Management and Controls, Business Model and Strategy, and Treasury and Asset Liability Management.

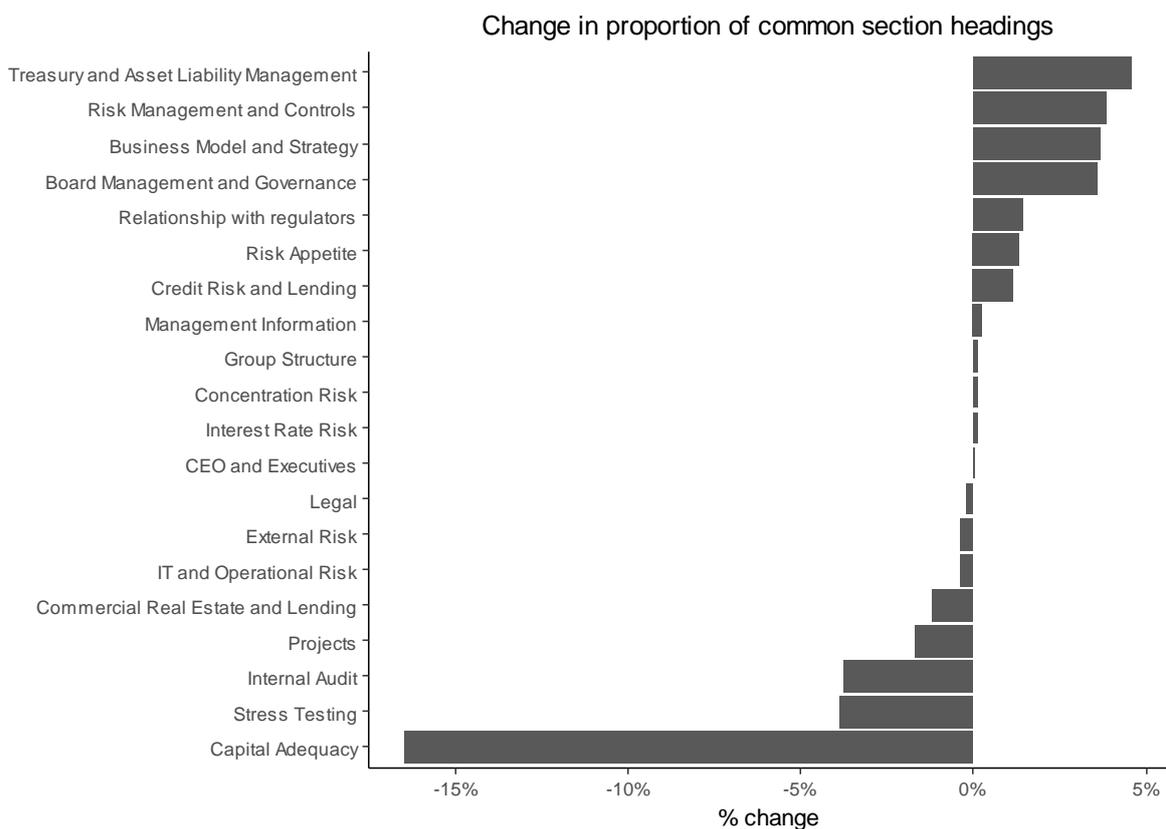


Figure 31: Graph showing changes in proportion of common section headings between ARROW and 2015 PSM letters

## Conclusion

This paper has quantitatively analysed the text of PSM letters to understand how supervisory communications to firms has changed over time, and how it differs depending on the nature of the firm with whom the PRA is communicating. We caution that our results may be imperfect gauges of PRA supervisory communication with firms. Nevertheless, we believe our findings, such as they are, should interest a wide set of stakeholders.

For the PRA, we find support that its supervisory communications are proportionate with respect to firm risk. Figure 32 summarises the main findings from section 4. Category 1 firms, which present greater inherent risk to UK financial system safety and soundness, and firms staged PIF 3-4, which present more imminent risk, are sent letters that are stylistically more complex and risk-focussed. Counterintuitively, these letters are less directive than those to less risky firms. One reason may be because the challenges facing such firms are complex, and therefore less amenable to direct, prescriptive instructions. Another possible explanation is that PRA supervisors have regular close and continuous meetings with the largest firms. The ongoing close relationship may enable supervisors to achieve their objectives with other oral or written communications in addition to the formal PSM letter process. Relatedly, we also find that the letters to these firms are also less formal, which could have a similar explanation.

Dimension of	High Risk Firms	Low Risk Firms	Supporting Linguistic Features
<b>Proportionality</b>	CAT1 or PIF3-4	CAT2-4 or PIF1-2	
<b>Complexity</b>	more complex	less complex	length of letter (in words) proportion of high risk diction (out of total number of words)
<b>Sentiment</b>	more negative	less negative	proportion of obligatives (out of total number of words)
<b>Directiveness</b>	less directive	more directive	proportion of deadlines (out of total number of words) proportion of 'please' (out of total number of words) ratio of sentence initial 'please' : sentence medial 'please'
<b>Formality</b>	less formal	more formal	typed vs. handwritten salutations generic vs. named salutations

Figure 32: Overview of results from section 4

Our results should also give the public confidence that banking supervision has changed since the crisis. Our model suggests letters written by the PRA are distinct from those written previously by the FSA. We find that a range of linguistic features discriminate between the two types of letters. For example, we find that future-oriented diction is one of the key linguistic discriminators. PRA letters are overwhelmingly forward-looking—in line with its aspirations. We also find that the PRA's letters are relatively more complex, directive and formal. In terms of content, we find that the PSM letters have far fewer generic section headings, which we interpret as meaning that the PRA communicates key risks to firms in a more detailed way that is tailored to their idiosyncratic risks.<sup>36</sup> We also found a strong focus on the topics of liquidity, and recovery and resolution planning in the PSM letters. These were absent in the ARROW letters. In this regard, PSM letters reflect the shift in the

<sup>36</sup> It may also reflect the increasing complexity of financial regulation since the crisis (Haldane and Madouros 2012).

supervisory agenda post financial crisis. Our findings thus resonate with other information about the effectiveness and quality of the PRA's approach to its relationship with firms. For instance, feedback solicited from firms by the PRA shows that during the 2016-2017 supervision year, 97% of Category 1 to 4 firms agreed that their firm has an effective relationship with the PRA. And 91% agreed that their firm is clear what the PRA's expectations are as to what it needs to do to address key risks.<sup>37</sup>

Finally, here are some ways the research we have conducted could be extended by other researchers.

- Although our focus was on banks and building societies, the same PSM process applies to insurance firms. One could analyse insurance PSM letters in a similar way.
- We have used random forests in this paper. This is a supervised machine learning algorithm, where input data (the letters) have been labelled. Alternatively, other researchers could use an unsupervised machine learning approach (Chakraborty and Joseph 2017). This might identify clusters of firm letters quite apart from their Category and PIF classification.
- The analysis could be extended to study changes over time. A line of research that may be worthwhile would investigate linguistic correlates of improvement/worsening of PIF scores for a firm.

---

<sup>37</sup> <http://www.bankofengland.co.uk/pradocuments/supervision/firmfeedback201617.pdf>

## References

- Bank of England. (2014). The Bank of England's Approach to Resolution. Available from <http://www.bankofengland.co.uk/financialstability/Documents/resolution/apr231014.pdf>
- Bank of England. (2016). The Prudential Regulation Authority's approach to banking supervision. Available from <http://www.bankofengland.co.uk/publications/Pages/other/pr/supervisoryapproach.aspx>
- Bholat, D & Gray, J. (2013) Organizational form as a source of systemic risk. *Economics* 7: 1-35.
- Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks. London: Bank of England.
- Biber, D. (1991). Variation across speech and writing. Cambridge: Cambridge University Press.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. Wadsworth, Belmont.
- Chakraborty, C & Joseph, A. 2017. Machine learning at central banks. London: Bank of England.
- Correa, R., Garud K., Londono J. M., & Misleng N. (2017). Sentiment in Central Banks' Financial Stability Reports. International Finance Discussion Papers 1203. Board of Governors of the Federal Reserve System (U.S.).
- Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In Proceedings of ACL.
- Davies, H. (2010). The Financial Crisis: Who's to blame? Cambridge: Polity Press.
- FCA and PRA. (2015). The failure of HBOS plc (HBOS). London: Bank of England.
- Financial Services Authority, Internal Audit Division. (2008). The supervision of Northern Rock: a lessons learned review. London: FSA.
- Financial Services Authority. (2009). The Turner Review: A regulatory response to the global banking crisis. London: FSA.
- Financial Services Authority. (2011). The failure of the Royal Bank of Scotland. London: FSA.
- Goldsmith-Pinkham, P., Hirtle, B., & Lucca, D. (2016). Parsing the Content of Bank Supervision. New York: Federal Reserve Bank of New York.
- Haldane, A. & Madouros, V. The dog and the frisbee. Speech at the Federal Reserve Bank of Kansas City's 36th economic policy symposium on 31 August 2012.
- Hansen, S, McMahon, M & Prat, A. (2014). Transparency and deliberation within the FOMC: a computational linguistics approach. CEPR Discussion Paper 9994.

- Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Berlin: Springer
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18),2543-2546.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin: Springer.
- Henry, E. (2006). Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting* 3, 1-19.
- Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication* 45, 363-407.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. Wiley, Hoboken.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3),651-674.
- House of Commons Treasury Committee. (2008). *The run on the Rock*. London: The House of the Commons.
- Huddleston, R. & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Janitza, S., Strobl, C., & Boulesteix, A.-L. (2013). An AUC-based permutation variable importance measure for random forests. *BMC bioinformatics*, 14(1), 119.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Allen Lane.
- Kane, E. (2015). A Theory of How and Why Central-Bank Culture Supports Predatory Risk-Taking at Megabanks. *Institute for New Economic Thinking Working Paper 34*.
- Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.
- Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.
- Merrouche, O., & Nier, E. (2010). *What Caused the Global Financial Crisis? - Evidence on the Drivers of Financial Imbalances 1999-2007*. International Monetary Fund.
- Pinker, S. (2014) *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*. London: Penguin.
- Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. Springer.
- PRA and FCA. (2016). *The PRA's and FCA's Threshold Conditions*. Available from: <http://www.bankofengland.co.uk/pradocuments/authorisations/newfirmauths/thresholdconditionsfactsheet.pdf>
- Provost, F. & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. Sebastopol: O'Reilly.

Saussure, F. de. (1983). *Course in General Linguistics*. Eds. Charles Bally and Albert Sechehaye. Trans. Roy Harris. La Salle, Illinois: Open Court.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1),307.

Strobl, C., Malley, J., & 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.

Turner, A. (2015). *Between Debt and the Devil: Money, Credit, and Fixing Global Finance*. Princeton: Princeton University Press.

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

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., & Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13),1636-1643.

## Annex 1: Data selection criteria

We started our analysis by establishing the population of firms by reference to the PRA Supervisory Approach Team's spreadsheet. This spreadsheet shows a comprehensive list of entities supervised by the PRA, along with the Category assigned to each entity. After identifying the population, we restricted our analysis to firms meeting the following criteria:

- Only firms supervised by the UK Deposit Takers Directorate. PRA supervision is the responsibility of Insurance Supervision, International Banks Supervision and UK Deposit Takers Supervision.<sup>38</sup> Our analysis focused only on firms regulated in the last of these directorates for reasons spelled out below.
- Only UK firms i.e. firms headquartered in the UK. The nature of the supervisory relationship between the PRA and firms, and thus the nature of communication, will differ depending on whether the PRA is the primary home regulator or a secondary host regulator. By construction, we included only firms where the PRA is the primary regulator, establishing a common basis for comparative purposes.
- Banks and building societies only. We excluded credit unions, investment banks, insurance firms and friendly societies. We excluded insurers and investment banks by construction because their regulation and business models are distinct from deposit taking UK banks. Credit unions and friendly societies were excluded as they are seen to have a low impact on the economy in the event of their failure.
- Category 1-4 firms only. All Category 5 firms were excluded as these firms are not systemically important and are not supervised in quite the same way as Category 1-4 firms.
- Consolidated firms, where applicable. Where a firm has both important UK and Group operations, it may receive two separate letters. We used only the consolidated entity's letter to avoid double counting.
- No newly authorised firms. We began analysis in May 2016. Firms which had been recently authorised and therefore were considered a 'Newly Authorised Firm' would not have had at least two Periodic Summary Meeting (PSM) letters. We required there to be at least two PSM letters per firm to boost our sample size.
- No firms which had undergone significant changes in control recently i.e. mergers or acquisitions. Letters pre and post such changes in control were likely to be too different from each other for comparative analysis.
- Firms not in administration. A firm is no longer seen as a going concern firm if it is in administration. In such cases, the normal rules of supervisory engagement do not apply. This excluded PIF stage 5 firms by construction.

---

<sup>38</sup> Insurance Supervision deals with general insurers and life insurers, friendly societies and London Markets firms (the Society of Lloyd's, its managing agents and general insurance and reinsurance companies operating in London market). International Banks Supervision comprises investment firms, international banks and custodians and is the host regulator of the UK activities of firms which are head-quartered in over 50 overseas jurisdictions. UK Deposit Takers Supervision is responsible for the supervision of banks, building societies and credit unions and is further split into Major UK Deposit Takers (the largest UK deposit takers) and the Banks, Building Societies and Credit Unions (smaller UK banks, building societies and credit unions).

## Annex 2: Linguistic features

### Complexity

We explored features of complexity at two levels of analysis—that of the document<sup>39</sup> and that of the sentence. Unless otherwise stated all document-level metrics are total counts within a letter (i.e. “rate per document”), while, for sentence-level metrics, raw counts are divided by the number of sentences within the document (i.e. “rate per sentence”).

#### *Document-Level Complexity Metrics*

**Word Count** We use the total number of words in the letter as a simple measure of complexity. Because of variation in how compound words are written—sometimes with a hyphen (e.g. *firm-specific*) and sometimes without (e.g. *firm specific*)—we treated the components of such words as distinct tokens; thus, this example counts as 2 words not 1.<sup>40</sup>

**Section Headings** The PSM letter is typically divided into sections, occasionally more than one-level deep. For example, in a letter to one firm, the section heading “Key Risk Areas” has a subsection (“Treasury Capability”) which itself has a subsection (“IRRBB”). Thus, we consider the number of sections in a letter to be an additional complexity feature. We manually extracted all section headings from each letter, regardless of their level of embedding, and counted them. In the example above, the count would be three headings.

**Appendix Presence**<sup>41</sup> We examined whether the presence of an appendix might correlate with a firm’s degree of risk. The presence of an appendix often means that a stand-alone piece of detailed assessment work has been done. The presence of an appendix may therefore indicate a firm with more complex risks and a higher degree of supervisory oversight.

**Acronyms** We included a feature for the number of acronyms (e.g. CCR, ICAAP) in a letter. We use acronyms in an informal way here to include acronyms proper (e.g. ICAAP), which are pronounced like ordinary words, and abbreviations (e.g. CCR), which are pronounced character-by-character. In the linguistics literature, abbreviations and acronyms are sub-types of initialisms (Huddleston and Pullum 2002: 1632–1634). The use of acronyms, particularly infrequent ones, increase a document’s complexity, as the reader has to decode them. To measure this feature, we compiled a list of 172 prudential policy acronyms from the PRA’s intranet page. The total number of these acronym tokens in a letter was divided by the total number of words to remove the effect of letter length.

**Numerals** We assume that an increase in the number of numerals in a document corresponds to an increase in complexity because numerals provide additional, specific detail. As humans typically have innate difficulties processing numbers (Kahneman 2011), more of them in a document make the text cognitively more demanding to decode.

---

<sup>39</sup>An obviously straightforward measure of document-level complexity is the total number of pages in a letter. However, we decided not to include it principally because the font and font sizes differed markedly between the letters. A better proxy is total number of words.

<sup>40</sup> Determining what constitutes a ‘word’ is not entirely straightforward. For an interesting discussion in the information retrieval literature, see Manning et al. (2008: 24–25).

<sup>41</sup> Some appendices are included as standard in PSM letters, for instance, Individual Capital Adequacy Assessment Process (ICAAP) appendices, and Individual Liquidity Adequacy Assessment Process (ILAAP) results.

## Sentence-Level Complexity Metrics

Complex sentences typically include explanations and other pieces of information which support the key, unadorned message of the sentence. One might surmise that letters with more sentential detail (and thus higher sentence complexity) will be sent to firms with a higher level of inherent and imminent risk because, for these firms, the scale and complexity of risk analysis will be greater. As such, the communication of this analysis needs to be more sententially complex. On this basis, we include several measures of sentence-level complexity in our feature catalogue.

**Sentence Length** To measure sentence length, we divided the total number of words in a letter by the number of sentences in the letter.

**Punctuation Rate** As a second measure of sentence complexity, and as a coarse proxy for the number of clauses in a sentence, we measured the punctuation rate in a sentence. This was computed by counting the number of punctuation marks (including periods) in a letter, and dividing this count by the letter's sentence count.

**Subordination Rate** As a third proxy for sentence complexity, we counted the mean number of indicators of clausal subordination per sentence. Clausal subordinators serve to highlight that one clause grammatically depends on a constituent in a higher clause. Consider the examples in (1), gleaned from the PSM letter corpus.

- (1) a. . . .we felt [that this point was worthy of feedback] . . .  
b. . . . provide the PRA with a summary of work [completed to date] on assessing ongoing IT resilience...

In (1-a), *this point was worthy of feedback* is embedded inside the *felt*-clause, and is grammatically signalled as such by *that*. In (1-b) the clause *completed to date* depends on the nominal *work* in the higher clause; the subordination in this example is signalled by the juxtaposition of a noun (*work*) and the past participle (*completed*).<sup>42</sup> For us, this variable provides a metric of how many clausal relationships are on average packed into a single sentence in a letter.

We operationalised this feature by using dictionary-based methods supplemented with Part-Of-Speech (POS) tagging where needed,<sup>43</sup> drawing on the following indicators of clausal subordination gleaned from the grammar books: (1) Subordinators – *although, because, if, though, unless, whereas, whereby, whereupon, while, whilst*; (2) WH-words – *who, whom, whose, which, where, when, whether, how, what, why*, including complex counterparts (*whoever, whomever, whichever, wherever, whenever, whatever*); (3) *That*; (4) *To*; (5) Subject-operator inversion – *Had, Should, Were*;<sup>44</sup> and (6) Reduced relatives (as in example (1-b)).

---

<sup>42</sup> This is an example of a Reduced Relative which could otherwise have been expressed as which/that has been completed to date.

<sup>43</sup> Part-Of-Speech (or POS) tagging assigns a word class label to a word, e.g. adverb, conjunction, etc.

<sup>44</sup> Subject-operator inversion can occur in counterfactual conditional sentences in place of the overt subordinator *if*, e.g. *If the firm had addressed this issue earlier it would not be a concern now* -> *Had the firm addressed this issue earlier it would not be a concern now*.

As POS-taggers do not discriminate between prepositions and subordinating conjunctions, we excluded words that could function as both (e.g. *after, as, before, since, until*), except for *to*. For *to*, we accepted sequences in which *to* was directly followed by a verb base form (i.e. the verb form with no ending) or followed by an adverb which was itself followed by a base verb form. Otherwise, we excluded it.

The word *that* has various functions in English grammar — it can function (i) as a “determiner”, that is either as a grammatical modifier of a noun, e.g. *that firm*, or as a pronominal determiner in place of a noun (e.g. *that is the reason...*), or (ii) as a subordinator to mark that an embedded clause follows, e.g. *we expect that the firm...*. As the former does not indicate subordination, we included only subordinator-tagged *that's* and excluded determiner-tagger *that's*.

The word *because* can serve as both a subordinator (as in (2)) or as a complex preposition with *of* (as in (3)). We thus excluded *because of* sequences.

- (2) <Firm> fails to adequately capture, manage and monitor its credit risks because it does not have the appropriate risk management framework in place.
- (3) The legacy commercial loan book remains a long-term risk to the Society's overall performance, because of its poor credit quality.

We counted the number of extracted subordinate clause indicators in a letter and divided by the total number of sentences in that letter.

**Verb Rate** As another proxy for sentence complexity, we counted the rate of verbs per sentence. The inclusion of this attribute is meant to capture the fact that subordinators may not be overt. For instance, compare the sentences in examples (4)–(5), below. The (a) sentences have an overt indicator of subordination, ‘that’ and ‘to’ respectively, whilst the (b) sentences do not.

- (4) a. We believe [that the society is hindered from doing so. . . ]  
b. We do however believe [the current forecasted plan is over-optimistic].
- (5) a. . . . could help [to mitigate these].  
b. . . . should help [boost capability].

To provide a measure for the rate of verbs per sentence, we counted the number of verbs in a letter and then divided this count by the total number of sentences in the document.

### Sentiment

To measure how the PRA expresses its (dis)satisfaction with firms in the letters, we explored a set of variables that measure the balance between negative and positive diction, and vocabulary associated with high risk.

**Sentiment Analysis** The sentiment of an entire letter can be quantified by giving it a sentiment ‘score.’ The more negative the score, the more negative the sentiment, and vice versa. The score is calculated by taking the difference in the number of ‘positive’ and ‘negative’ words and normalising (dividing) by the total number of words in the letter.

‘Positive’ and ‘negative’ words refer to words that convey positive and negative sentiments, respectively, and can be found by searching for matches between the words in the letter and a dictionary. General sentiment dictionaries have been shown to perform poorly on financial texts, because words which ordinarily have negative connotations (e.g. *liability* and *tax*) are in fact quite neutral in a financial context (Loughran and McDonald, 2011).<sup>45</sup> We therefore used a finance-specific dictionary:

([http://www3.nd.edu/~mcdonald/Word\\_Lists.html](http://www3.nd.edu/~mcdonald/Word_Lists.html)).

**High Risk Diction** We qualitatively examined the letters and other documentation gleaned from the PRA’s intranet pages to establish a list of vocabulary items the PRA commonly uses when discussing high risk firms. Upon examining this material, we identified three main types of diction. First, there is the use of the term “high risk” itself. For this, we counted how often this phrase was used in a letter, including synonymous phrasing e.g., *high-risk*, *high level of risk*.

Second, we identified diction relating to ‘How’ and ‘Why’ a firm might pose a risk to the PRA’s statutory objectives. More specifically, such diction relates to a firm’s (a) Vulnerabilities (*exposed*, *fragile*, *susceptible*, *vulnerable*, *weak*); (b) Difficulties (*block*, *delay*, *difficulty*, *disrupt*, *encumber*, *hamper*, *hinder*, *hindrance*, *impediment*, *obstacle*, *problem*, *slow*, *stop*); (c) Doing too little of the right thing (*absence*, *deficient*, *inadequate*, *incomplete*, *insufficient*, *lack*, *shortcoming*, *too few*, *too little*); and (d) Doing too much of the wrong thing (*disproportionate*, *excess*, *extreme*, *over-compounds*, *overly*, *too many*, *too much*, *unattainable*, *undue*, *unfeasible*, *unnecessary*, *unreasonable*).

The third vocabulary set we constructed relates to how the PRA will monitor the risk. We counted the number of explicit references to the PRA’s Watchlist, as well as vocabulary that the PRA uses when communicating with high-risk firms, and reflect increased supervisory oversight – *check*, *closely*, *keep track of*, *monitor*, *oversee*, *scrutiny*, *scrutinise*.

To provide a single index, we summed these counts and divided by the total number of words in the letter.

### Directiveness/Directness

For this dimension, we considered directive expressions in the letters and the degree of politeness with which such directives were expressed.

**Obligation** We examined how frequently the PRA obliges a firm to perform some action: ‘Firm must do X.’ We coded for a single feature relating to various ways of expressing obligation. First, we included strongly obligative modal auxiliaries, as in the examples in (7) (taken from the PSM letters).<sup>46</sup>

---

<sup>45</sup> See also Henry (2006, 2008) and Correa et al. (2017).

<sup>46</sup> Aside from their obligative (“deontic”) semantics, these auxiliaries can convey other meanings. For instance, ‘should’ also can have an epistemic connotation which conveys the speaker’s knowledge about some state of affairs, as in (6).

(6) We consider the roll-out of <Initiative> should help boost capability.

In this example, the PRA is not issuing an order, but rather giving an indication of its knowledge concerning whether or not the initiative mentioned will help increase capability. However, as Huddleston and Pullum

- (7) a. The risk function must be reviewed and strengthened.  
 b. <Firm> should fully consider the implications of this decision.  
 c. There remain issues that ought to be addressed.

Second, we included a set of verbal constructions indicating strong modality. Specifically, we counted the frequency of the string [*PRA/we* (Adverb) (*will/would*) (Adverb)] followed by any of the following verbs:<sup>47</sup> *ask(s)*, *expect(s)*, *need(s)*, *request(s)*, *require(s)*, *want(s)*.<sup>48</sup> We give some examples from the letters in (8).

- (8) a. We expect the board to continue to monitor progress.  
 b. . . . the specific actions that we ask the society to take.  
 c. . . .we request that senior management provide us with. . .  
 d. . . .we require you to notify us if the ratio exceeds 35%.

Thirdly, we included adjectival constructions indicating strong obligative modality. For this, we counted the frequency of *it is* followed by any of the following adjectives: *critical*, *crucial*, *essential*, *imperative*, *important*, *necessary*, *vital*. (9) provides some examples from our corpus.

- (9) a. . . . it is imperative that the society performs due diligence. . .  
 b. . . . it is essential that it manages its mortgage book closely.  
 c. . . . it is crucial that <Firm> has an effective recovery plan.  
 d. . . . it is vital the board ensures momentum is maintained. . .

We summed the counts and divided the sum by the total number of words in a letter to remove length effects.

**Deadlines** We presume that if a request is given a deadline it is more pointed than one that is not. We surmise that, if there is no deadline, a firm may feel less compulsion to oblige with the 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. This count was then normalised by dividing by the total number of words in the letter.

**Politeness Indicators** One would assume that the most impolite language (that is, the most direct and blunt language) is used when communicating with firms that pose substantial risks because it is imperative for those firms to address issues with urgency. Too much politeness may mean the firm feels less compelled to comply, as the PRA may be seen as being “too nice.” We explored the following indicators.

**Please** As a basic marker of politeness, we counted the number of times *please* occurs in each letter, and divided this count by the total number of words in the letter. Additionally, Danescu-Niculescu-Mizil and co-authors (2013) find politeness differences with respect to

---

(2002) point out, this epistemic connotation occurs much less frequently than the deontic sense. We do not think it biases the results if we include all counts for these modals.

<sup>47</sup> Restricting the search to this particular syntactic frame may seem overly-restrictive. However, it was important to do this, in order to avoid counting sequences such as *you have indicated that the board of <FIRM> expects to make a final decision. . .*, etc.

<sup>48</sup> Other verbs are used in a similar context, e.g. *suggest* and *recommend*, but they are only weakly deontic as they do not constitute orders as such. Accordingly, we exclude them.

the position of *please* in a sentence—with sentence-initial *please* being impolite/direct (as in (10)) and sentence-medial *please* being polite/indirect (as in (11)).

(10) Please could you send us a copy of the corporate plan. . .

(11) Additionally, could you please also assess. . .

Accordingly, we counted the frequency of initial-*please* and the frequency of medial-*please*, and divided the two counts to provide an initial-*please*/medial-*please* ratio.<sup>49</sup> A higher value indicates more directness.

**Initial you** Danescu-Niculescu-Mizil et al. (2013) also find that sentence-initial second person pronouns, as in (12), have the same effect as initial *please*—they are apparently more direct than those that are concealed in sentence-medial position, as in (13). We therefore coded for an initial-*you*/medial-*you* ratio. Again, higher values are indicative of more direct phrasing.

(12) You should refer to this report. . .

(13) . . .we are willing to work with you on this journey . . .

### Formality

**Involvement** Involvement concerns the extent to which a speaker/writer involves themselves (e.g. through the explicit use of first person pronouns) and the listener/reader (e.g. through the explicit use of second person pronouns) in their speech/writing. In terms of genre distinctions, less formal and more personal narratives are typically more “involved” than formal and impersonal styles which are by contrast more “detached” (on the notion of involved styles, see Biber 1991: 43). To provide a general measure for this, we counted the total number of first or second (collectively termed “local”) person pronouns in a letter and divided by total word count, to eradicate letter length effects. We also examined some specifics of involvement, which we now discuss.

**Sender Self-Reference** We examined how the sender refers to themselves in the letters: (i) as a third-person collective body (“the PRA”), (ii) as a first-person collective body (“we”); or (iii) as a first-person individual (“I”). On the involvement–detachment continuum, (i) is more detached than (ii), whilst (ii) is more detached than (iii).

We separately collected by-letter frequencies for *PRA* (and -’s form *PRA*’s), first-person plural pronouns (i.e. *we*, *us*, *our*, *ourselves*), and first-person singular pronouns (i.e. *I*, *me*, *my*, *mine*, *myself*). From these counts, we defined three features:<sup>50</sup>

- *I/PRA* ratio
- *we/PRA* ratio
- *I/we* ratio

If the numerator is large relative to the denominator, this indicates a more involved (or in other words, more informal) style of writing.

---

<sup>49</sup> Note that we added a count of 1 to the numerator and denominator to avoid undefined values.

<sup>50</sup> Again, we added a count of 1 to the numerator and the denominator.

**Recipient Reference** We also examined how the PRA refers to the firm in the letters: (i) as a third-person collective body (e.g. “the firm”, “the society”, “the board”); (ii) as a second-person collective body or individual (“you”).<sup>51</sup> On the involvement–detachment scale, (i) is more detached than (ii).

For (i), we manually gathered a list of company names of the firms (usually an initialism), and used pattern matching to count their by-document frequency. We added to this count the frequency of two generic ways of referring to the firm – the phrase *the firm* (used for banks) and the phrase *the society* (used for building societies) – and the number of mentions of the phrase *the Board* (hits for *the Board of the PRA* and the like were excluded).

For (ii), we counted all forms of the second-person pronoun (i.e. *you*, *your*, *yours*, *yourself*, *yourselves*). Phrases such as *your board*, *your firm* and so on were included in this count because they explicitly index the second person by means of *your*.

From these counts, we defined a *you/firm* ratio feature, where a larger number indicates increased involvement (i.e. more personalness and less formality).

**Style of Salutation** 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 (e.g. “*Dear Geoffrey*”) is obviously much less formal than one that is addressed to a generic collective of individuals (e.g. “*Dear Sirs*”). Further, a letter that has a handwritten salutation is less formal than one that is typed. We coded for these two features: (i) whether the letter has a generic address formulation, e.g. *Board of Directors*, *Board Members*, *Directors*, *Sirs*, *Sirs and Madam* or to a named individual; and (ii) whether the letter has a typed address formulation or a handwritten one.

### Forward-lookingness

**Tense** We examined the grammatical tense composition of the letters. The grammar of English has two tenses, Past, as in “the firm breached its internal tolerances”, which typically relates to past time, and Non-past, as in “it is essential that it manages its mortgage book closely”, which typically relates to present or future time.<sup>52</sup> In addition, while there are not strictly tenses, English has a Perfect construction (e.g., “*We note the progress that <Firm> has made*”) where a past event has current relevance; a set of Futurates (e.g., “*is to be. . .*”, “*be about to. . .*”, “*be going to. . .*”) which reference future time; and a set of Modals (e.g., *can*, *could*, *may*, *might*, *ought*, *shall*, *should*, *will*) which have a future ‘feel’ to them. In this analysis, we grouped together Past tensed verbs and Perfect constructions as Past-oriented tense, and grouped together Non-past, Futurates and Modals as Non-past-oriented tense—where “tense” should be interpreted loosely and not in the strict grammatical sense. To form our metric, we calculated the percentage of verbs with Non-past-oriented tense out of total number of verbs that exhibited tense marking in the letters, i.e. the sum of the counts for Past-oriented verbs and Non-past-oriented verbs.

**Future-Oriented Diction** While our first forward-looking metric is grammatical in nature, our second is lexical, i.e. words with semantic content.<sup>53</sup> For the latter, we examined the

---

<sup>51</sup> We do not make a distinction between collective/plural *you* and individual/singular *you* because their word forms are identical except for the reflexive *yourself*/*-selves*.

<sup>52</sup> Note that English doesn’t have a distinct future tense (Huddleston and Pullum 2002: 208–210).

<sup>53</sup> In linguistics, a distinction is made between grammatical (or functional) elements and lexical (semantically contentful) elements. To clarify this distinction, consider the sentence *John will arrive*. The words *John* and *arrive* provide the real meaning, while *will* provides ancillary functional information that relates an event (*John’s arriving*) to time. As another example, consider *John arrived*. Here, again *John* and *arrive* provide the

percentage of sentences in a letter containing a “future-oriented” word. To establish a list of future-oriented words, we drew upon and augmented the lists suggested by Li (2010) and Bozanic et al. (2013), who investigated forward-looking statements in corporate filings. Our list of future-oriented words are: *ahead, aim, aims, aiming, anticipate, anticipates, anticipating, anticipation, believe, believes, belief, approaching, coming, continue, continues, ensuing, estimate, estimates, expect, expects, expectation, forecast, forecasts, forecasting, forthcoming, forward, future, goal, goals, hope, hopes, hoping, impending, imminent, incoming, intend, intends, intending, intention, later, next, objective, outgoing, outlook, plan, plans, planning, potential, potentially, predict, predicts, predicting, prediction, projecting, projection, prospect, prospects, prospective, schedule, schedules, succeed, succeeds, succeeding, succession, target, targets.*

---

content, while the *-ed* suffix on *arrive* adds functional information concerning the relative time of John’s arrival (i.e. that it happened before the moment of speaking). In other words, functional elements help to give a sentence its formal structure, while lexical elements provide information that adds meaning.

---

### Annex 3: Random forest methodology

In this analysis we make use of the machine learning algorithm of random forests to predict our binary response variables based on the suite of linguistic features mentioned in Annex 2. Random forests, in their original implementation, are based on a Classification and Regression Tree algorithm (CART) developed by Breiman and co-authors (1984). The algorithm takes the full dataset, identifies a predictor with a split-point that is most influential at distinguishing classes, and splits the dataset into two parts. The procedure then works through these partitions, and continually sub-divides them until a stopping criterion, such as the minimum number of observations in a particular node, is reached.<sup>54</sup>

The original CART algorithm has been shown to be biased in terms of how it selects features at each node, with quantitative features and qualitative features with many levels preferentially chosen over binary features (Hothorn et al. 2006). This is because the CART algorithm considers all split points amongst all features simultaneously, and thus those with a larger number of potential split-points (i.e. quantitative features) have a greater chance of being chosen than those which only have a single split-point (e.g. binary categorical features). To address this problem, Hothorn and co-authors (2006) developed a method in which variable selection and splitting are considered as two separate stages in the decision-making process, rather than being considered simultaneously. In their approach, a test of significance is performed on each feature to determine whether it is significantly related to the response variable. The feature with the strongest association with the response variable is subsequently probed in more detail to locate that split-point which separates the response classes the best. This is implemented in their conditional inference tree (ctree) algorithm, which we use to grow our forest.

A major problem with individual trees is that they are sensitive to minor changes in the original data. Specifically, different trees grown on different subsamples of the data can yield trees that are very different in their composition. As a result, this means that a single tree typically overfits to the data, making them poorly suited to prediction on new data. Instead of basing prediction on a single tree, Breiman (2001) developed a random forest algorithm which bases prediction on an ensemble of trees. Here, a large number of trees (ntree) are grown, each using a different bootstrap sample or subsample from the original dataset.<sup>55</sup>

Besides random subsampling, random forests have an additional facet of randomness in their design, in that at each node only a random subset of the full set of features is evaluated for splitting. Restricting the candidate feature set in this way is important in terms of out-of-sample predictive accuracy. If all features are available for evaluation, then dominant features will repeatedly show up in the earliest nodes in the trees, making the trees in the forest and thus their predictions somewhat correlated. When the resulting ensemble classifier is tested on out-of-sample data, it may not perform well enough simply because those dominant features do not pertain as strongly to the new data. However, if the dominant feature cannot be chosen because it is not randomly selected as one of the features to be evaluated, then

---

<sup>54</sup> We will not go into the specifics of how a split-point is determined to be influential, nor the exact nature of the stopping criterion, as these differ depending on which tree algorithm is used. Instead, we refer the reader to Strobl et al. (2009) for a general discussion, and to Hothorn et al. (2006) for details.

<sup>55</sup> Bootstrapping refers to randomly sampling  $n$  observations with replacement from a dataset of size  $n$ , such that a total of  $0.632 \times n$  observations make their way into the bootstrap. Subsampling refers to sampling  $0.632 \times n$  observations without replacement from a dataset of size  $n$ . For the reasoning behind the 0.632 factor, see Hastie et al. (2009). In our implementation, we use subsamples as that is the default in the R function we use. For the theoretical motivation behind using it instead of bootstrapping see Politis et al. (1999).

other weaker predictors, which may be more prominent in the new data, get the chance to manifest their effect and interact with other variables. As a result, the trees that make up the forest are much more diverse and their “average” may more adequately capture population effects than an “average” biased towards an in-sample dominant feature. For classification problems, the size of this subset (termed *mtry*) is generally recommended to be the square root of the total number of features.<sup>56</sup> In our analysis, we use the recommended *mtry* setting, which for our datasets is  $\sqrt{25} = 5$ .

**Performance** In machine learning approaches, it is usual to divide the original dataset into at least two parts –one larger set used to train the classifier, and a smaller set used to test how generalisable the model is when it comes to unseen data (see Provost and Fawcett (2013) for a discussion of the logic behind this approach). In random forests, training and testing is done “internally.” Each tree is trained on a different random subsample of 63.2% of the observations from the original dataset (these are termed “in-bag” observations). This leaves 36.8% of the data (termed “out-of-bag” or OOB observations) which are used for testing the predictive ability internally within the model. Specifically, each tree in which an observation is OOB is used to determine the response category that that observation is most likely to belong to, based on the values it assumes for the predictor variables. The majority decision from all the trees in which that observation is OOB is then taken as the “winning” category for that observation.

To clarify with an example: let us assume we have built a random forest model consisting of 10 trees to predict whether a firm is a bank or a building society (b. soc) based on a set of variables in a dataset consisting of 16 observations. For each tree that is grown, ( $0.632 \times 16 \approx$ ) 10 randomly chosen observations are used for training, with the remaining 6 used for testing. We give a tabular depiction of this in Figure A3-1. The first column indicates the observation case number. The second column shows the observed outcome—‘bank’ or ‘b. soc.’ The columns T1 through T10 indicate the ten trees in the forest. In each of these columns, the 10 cells labelled ‘-’ indicate that a particular observation was used for training the tree. The remaining 6 labelled cells indicate that the observation was OOB along with the value the tree returned as the most likely class for that observation—‘bank’ or ‘b. soc.’ The final column ‘Winning Class’ shows the majority label. For example, for the first observation, ‘bank’ is the majority decision, returned in 3 out of the 4 trees in which this observation was OOB.

---

<sup>56</sup> This is a tuning parameter. Its optimal value may be sought through a random grid search. One might consider evaluating a range of *mtry* values from, say, 2 to *p* on a subset of the original data (the validation set) and use that *mtry* value that optimises performance for training. We simply use the recommended *mtry* value, with the caveat that it may not be the optimal one in terms of performance.

Observation	Outcome	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Winning Class
1	bank	-	bank	-	-	bank	-	-	bank	b. soc.	-	bank
2	bank	b. soc.	-	-	b. soc.	-	-	-	-	bank	b. soc.	b. soc.
3	bank	-	bank	-	bank	bank	-	bank	-	-	-	bank
4	bank	-	-	bank	-	-	-	-	-	-	bank	bank
5	bank	-	-	b. soc.	b. soc.	-	b. soc.	bank	b. soc.	-	b. soc.	b. soc.
6	bank	-	-	-	-	-	bank	-	b. soc.	-	bank	bank
7	bank	bank	-	bank	-	-	-	-	-	-	-	bank
8	bank	-	bank	b. soc.	-	bank	-	bank	-	b. soc.	-	bank
9	b. soc.	-	-	b. soc.	-	b. soc.	-	-	b. soc.	b. soc.	b. soc.	b. soc.
10	b. soc.	b. soc.	-	bank	bank	b. soc.	b. soc.	-	b. soc.	-	-	b. soc.
11	b. soc.	-	b. soc.	-	-	b. soc.	b. soc.	-	-	-	b. soc.	b. soc.
12	b. soc.	bank	bank		b. soc.	-	-	bank	bank	-	-	bank
13	b. soc.	-	-	-	-	-	b. soc.	-	-	-	-	b. soc.
14	b. soc.	b. soc.	-	-	-	-	-	b. soc.	-	-	-	b. soc.
15	b. soc.	bank	-	-	-	-	-	bank	-	b. soc.	-	bank
16	b. soc.	-	bank	-	b. soc.	-	b. soc.	-	-	b. soc.	-	b. soc.

Figure A3-1: Example showing the prediction output of a toy random forest model

The forest’s OOB predictive accuracy can then be computed by reference to a so-called “confusion matrix”— a cross-tabulation of the winning class with the observed outcome in the data sample. Figure A3-2 illustrates using our hypothetical example. Here, 6 ‘b. soc.’ outcomes were correctly predicted as being ‘b. soc.’ (top-left cell), while 2 ‘b. soc.’ outcomes were wrongly predicted as being ‘bank’ (top-right cell). Similarly, 6 ‘bank’ outcomes were correctly predicted as being ‘bank’ (bottom-right cell), while 2 ‘bank’ outcomes were wrongly predicted as being ‘b. soc.’ (bottom-left cell). To provide a basic measure of predictive accuracy, one could calculate the proportion of correct classifications out of the total number of classifications made. Thus, we sum over the leading diagonal (in this example  $6 + 6 = 12$ ) and divide this sum by the total cell count ( $6 + 2 + 2 + 6 = 16$ ). In this example, OOB predictive accuracy is  $12/16 = 0.75$ , meaning that the classifier is correct 75% of the time.<sup>57</sup> This is an improvement upon the predictive accuracy of a pure random guess (50%).

		Winning Class	
		b. soc.	bank
observed outcome	b. soc.	6	2
	bank	2	6

Figure A3-2: Example showing the confusion matrix of the toy random forest model. Correct decisions are colour-coded in green.

<sup>57</sup> For the pessimists, this translates into an OOB error rate of 0.25.

In datasets with severe response category imbalances, it is relatively easy for the classifier to obtain a very high accuracy simply by guessing the empirically most frequent category all the time. For instance, in the Category dataset, one could obtain an accuracy of 93% by always predicting ‘Category 2–4’. As such, this straightforward type of performance metric can be quite misleading (for an extensive discussion of this issue, see Provost and Fawcett 2013). In the present paper, we instead make use of Harrell’s index of concordance  $C$  (Harrell et al., 1982), which is more robust against data imbalances.

$C$  makes use of OOB class probabilities, i.e.  $P(y_i = k)$  where  $y_i$  is an observational unit and  $k$  is the response class of interest (e.g. ‘bank’ in our example).<sup>58</sup> As a result, each observation has a probability score for a specified class of interest (e.g. ‘bank’). Figure A3-3 illustrates this. The first column shows the observation number. The second column shows the observed outcome. The third column gives the model’s predicted probability that the observation belongs to the ‘bank’ category.

Observation	Observed Outcome	$P(y_i = \text{bank})$
1	bank	0.68
2	bank	0.4
3	bank	0.79
4	bank	0.81
5	bank	0.32
6	bank	0.66
7	bank	0.91
8	bank	0.6
9	b. soc.	0.35
10	b. soc.	0.4
11	b. soc.	0.78
12	b. soc.	0.47
13	b. soc.	0.31
14	b. soc.	0.2
15	b. soc.	0.6
16	b. soc.	0.22

Figure A3-3: Example showing the predicted probability output of the toy random forest model

Conceptually,  $C$  is the probability that a randomly chosen observation from the ‘bank’ category will exhibit a higher predicted probability of belonging to that class compared to a randomly chosen observation from the ‘b.soc.’ category. To calculate  $C$ , one takes all possible pairs of observational units where the first element of each pair is the predicted probability of a ‘bank’ outcome for a ‘bank’ observation (let us call this score  $i$ ) and the second element of each pair is the predicted probability of a ‘bank’ outcome for a ‘b. soc.’ observation (we’ll call this score  $j$ ).

<sup>58</sup> The computation follows the logic as for discrete class responses, i.e. each tree in which an observation is OOB returns a probability score for that observation belonging to a particular class. Then all the OOB probability scores for that observation are averaged.

- If  $i > j$ , we give that pair a score of 1 to indicate that it is ‘concordant.’ For instance, if for a selected pair, the predicted probability of a ‘bank’ outcome for a ‘bank’ observation is 0.68, as in row 1 in Figure A3-3, and the predicted probability of a ‘bank’ outcome for a ‘b. soc’ observation is 0.35, as in row 9 in Figure A3-3, then the probabilities are concordant with the observations because the probability score is higher for the former than it is for the latter.
- If  $i < j$ , the pair receives a score of 0 to indicate that it is ‘discordant.’ For instance, if for a selected pair, the predicted probability of a ‘bank’ outcome for a ‘bank’ observation is 0.66, as in row 6 in Figure A3-3, and the predicted probability of a ‘bank’ outcome for a ‘b. soc’ observation is 0.78, as in row 11 in Figure A3-3, then the probabilities are discordant with the observations because the probability score is lower for the former than it is for the latter.
- If  $i = j$ , the pair receives a score of 0.5. For instance, if for a selected pair the predicted probability of a ‘bank’ outcome for a ‘bank’ observation is 0.4, as in row 2 in Figure A3-3, and the predicted probability of a ‘bank’ outcome for a ‘b. soc’ observation is also 0.4, as in row 10 in Figure A3-3, then the probabilities are neither concordant or discordant with the observed value because the probability score are equal.

Each unique pair is evaluated according to the above: in Figure A3-4 we give the respective probabilities and the scores (1, 0, or 0.5) for our toy dataset. The resulting scores are summed and divided by the total number of possible pairs to give a value for C. In the present example, this works out as  $C = 52/64 = 0.8125$ . C lies between 0.5 (indicating a model that discriminates no better than chance) and 1 (indicating a model that discriminates between classes perfectly), with models with  $C > 0.8$  manifesting excellent discriminability (Hosmer et al., 2013, 177).

<i>i</i>	<i>j</i>	score									
0.68	0.35	1	0.68	0.78	0	0.68	0.31	1	0.68	0.6	1
0.4	0.35	1	0.4	0.78	0	0.4	0.31	1	0.4	0.6	0
0.79	0.35	1	0.79	0.78	1	0.79	0.31	1	0.79	0.6	1
0.81	0.35	1	0.81	0.78	1	0.81	0.31	1	0.81	0.6	1
0.32	0.35	0	0.32	0.78	0	0.32	0.31	1	0.32	0.6	0
0.66	0.35	1	0.66	0.78	0	0.66	0.31	1	0.66	0.6	1
0.91	0.35	1	0.91	0.78	1	0.91	0.31	1	0.91	0.6	1
0.6	0.35	1	0.6	0.78	0	0.6	0.31	1	0.6	0.6	0.5
0.68	0.4	1	0.68	0.47	1	0.68	0.2	1	0.68	0.22	1
0.4	0.4	0.5	0.4	0.47	0	0.4	0.2	1	0.4	0.22	1
0.79	0.4	1	0.79	0.47	1	0.79	0.2	1	0.79	0.22	1
0.81	0.4	1	0.81	0.47	1	0.81	0.2	1	0.81	0.22	1
0.32	0.4	0	0.32	0.47	0	0.32	0.2	1	0.32	0.22	1
0.66	0.4	1	0.66	0.47	1	0.66	0.2	1	0.66	0.22	1
0.91	0.4	1	0.91	0.47	1	0.91	0.2	1	0.91	0.22	1
0.6	0.4	1	0.6	0.47	1	0.6	0.2	1	0.6	0.22	1

Figure A3-4: Probability pairs along with individual concordance scores for the toy dataset

**Variable Importance** One advantage of using random forests is that it provides an in-built measure for quantifying how influential a variable is at predicting the response, no matter how many variables there are in the model. There are several ways to do this. The most popular is a “permutation” based method which evaluates a variable’s “mean decrease in accuracy.” In this approach, the OOB error rate ( $1 - \text{predictive accuracy}$ ) of each tree in the forest is denoted by  $ER_t$ , where  $ER$  is the ‘error rate,’ and  $t$  is an individual tree. For a predictor  $j$  whose variable importance we seek to calculate, its original values are randomly changed (“permuted”) in order to nullify its original relationship with response. Using the permuted predictor of interest, together with the remaining (‘unpermuted’) predictors, the OOB error rate is again recorded, denoted by  $ER_{tj}$ . The difference between the ‘before permutation’ and ‘after permutation’ error rates is calculated for each tree in the forest, the differences for all trees are summed, and the resulting sum is divided by the number of trees in the forest:

$$VI_j = \text{MeanDecreaseInAccuracy}_j = \frac{\sum_{t=1}^{ntree} (ER_{tj} - ER_t)}{ntree}$$

If the model’s performance decreases compared to when the predictor is unpermuted, this indicates that the predictor is useful. If performance worsens or remains unchanged, the variable is unimportant. In our implementation, we use a modified variable importance procedure that takes into account correlated predictors (Strobl et al. 2008) and class imbalance (Janitza et al. 2013).

**Dependency Plots** In order to gain insight into how our features are related to the target variables of interest, we provide dependency plots. These express how the model’s predicted probability of a letter belonging to a given Category (e.g. Category 1 vs. Category 2–4) or PIF Score (e.g. PIF 3–4 vs. PIF 1–2) changes at varying values of a given variable, while holding the values of the other features at their median values (for quantitative variables) or mode level (for qualitative variables).<sup>59</sup> These plots allow us to gauge how the feature provides a contribution to predictive accuracy, or in other words, the way in which a linguistic feature relates to letter type, holding other variables constant.

**Implementation** It is usual to grow a single forest. However, as random forests use different subsamples of the original data, different runs can yield slightly different performance estimates and, more importantly for us, different variable importance rankings. Consequently, it is possible that an atypical forest may overstate the importance of a particular feature simply by accident. In order to take such fluctuations into account, we grew 100 forests, with each forest comprising 2000 trees, and averaged the results.<sup>60</sup>

<sup>59</sup> These plots are a simplified version of partial dependency plots (see Hastie et al. 2009: 369ff.), which in a classification framework show the predicted probability (or log odds) of a given class at varying values of a feature  $x$  while averaging out the effects of the other features.

<sup>60</sup> Another study that uses 100 forest iterations instead of relying on a single forest is that of Vahlne (2017).

## Annex 4: Section heading analysis

