Comparing Manual Against Automated Measures of the Tone of Interim Management Statements: A Market Based Approach

Sheehan Rahman, Martin Walker and Thomas Schleicher[1]

30 November 2019

[1] We are grateful for comments from Nathan Joseph, Niamh Brennan, Edward Lee, Niamh Brennan, Norman Strong and Andy Stark. Schleicher and Walker gratefully acknowledge research support from the ESRC (grant number ES/J012394/1 and ES/K002155/1) and the ICAEW. We are also thank Mahmoud El-Haj of Lancaster University for providing access to linguistic analysis software.

Sheehan Rahman (corresponding author) is Lecturer in Accounting, Department of Economics and Finance, Brunel University London, UK. Refer all correspondence to sheehan.rahman@brunel.ac.uk.

Martin Walker is Emeritus Professor of Finance and Accounting, Alliance Manchester Business School, The University of Manchester, UK. Email: martin.walker@manchester.ac.uk.

Thomas Schleicher is Lecturer in Accounting & Finance, Alliance Manchester Business School, University of Manchester, UK. Email: thomas.schleicher@manchester.ac.uk.

Comparing Manual Against Automated Measures of the Tone of Interim Management Statements: A Market Based Approach

Abstract: This paper contributes to the debate on content analysis methods by comparing the linguistic tone of financial disclosures derived from manual content analysis with that from automated, that is, computer-assisted, content analysis. Using a sample of 1,022 UK Interim Management Statements (IMSs) we provide evidence that, compared to tone scored by automated wordlists, tone scored manually has greater explanatory power for abnormal stock returns around the IMS disclosure events. When net tone is replaced by separate measures of positive and negative tone, we find that the explanatory power of tone for the market response materially improves for the automated measures but not for the manual measure. Detailed comparisons of the manual and automated scores reveals specific limitations of the automated approaches.

Keywords: Disclosure Tone, Interim Management Statement, Content Analysis, Share Price Reaction.

EFMA Classification Codes: 200 [Market Based Accounting Research], 710 [Accounting Issues], 760 [Methodological Issues]

1. Introduction

The analysis of the linguistic tone, or sentiment, of financial narratives is an area of growing interest to academics (Henry & Leone, 2016; Loughran & McDonald, 2016). Disclosure tone, commonly referred to as a disclosure's positivity or negativity, is measured by either manual or automated textual analysis techniques (Henry & Leone, 2016). Although early textual analysis used manual techniques for computing tone, in recent years, computer-assisted word-counts have become the norm (Henry, 2006, 2008). A recent study by Henry and Leone (2016) employs automated analysis to compare earnings press release tone under four widely used wordlists. They find that wordlists specialized for financial communication have greater model explanatory power for short-window announcement-period abnormal stock returns than tone from non-specialised wordlists. A natural follow-up question is whether the tone computed by manual textual analysis has greater model explanatory power than tone computed with specialised financial communication wordlists.

Manual and automated textual analysis provide different sets of advantages and disadvantages for tone measurement. On the one hand, while the wordlist approach can process large volumes of disclosures in a fast, cheap and consistent manner, manual analysis is able to capture the contextual meaning of the words used in written text. On the other hand, while manual analysis is time consuming and costly to implement, automated analysis cannot capture textual subtleties employed in disclosures. As such, prior research has sometimes argued, without direct empirical evidence, that manual analysis is likely to provide a more accurate measure of the tone than automated wordlists due to its ability in capturing context (Clatworthy & Jones, 2003; Schleicher & Walker, 2010). In this paper, we address the aforementioned research gap by comparing the tone measured by manual and automated methods.

In particular, we compare the alignment of manual and automated measures of tone with abnormal stock returns around the release of UK Interim Management Statements (IMSs). IMSs provide trading and financial performance updates for the first and third quarters of the financial year by firms listed in EU regulated markets. They are short disclosures, typically one or two pages long, almost entirely comprised of textual narratives that allow managers full discretion over the content.

We use a sample of 1,022 IMSs of FTSE All-Share Index firms during 2008 – 2013. We measure the full-document IMS tone, alternatively by manual analysis and by using two widely used automated wordlists – the Henry (2008) wordlist and the Loughran and McDonald (LM) (2011) wordlists. These two wordlists, specialised for financial communication, have greater explanatory power for abnormal stock return than non-domain specific wordlists (Henry & Leone, 2016). In a preliminary analysis, we find that the average IMS tone based on all three measures are positive and that all three tone measures have a positive correlation with market returns, consistent with prior research (e.g. Henry, 2008; Henry & Leone, 2016; Loughran & McDonald, 2011).

For our main analysis, we adopt a short-window event study methodology, following Henry and Leone (2016), and compare the explanatory power of manual or automated measures of net tone for abnormal returns. We find that the manual model has significantly greater explanatory power than the automated models. As an additional analysis, we replace the net tone measure with separate measures for the positive and negative components of the tone and find that, while the explanatory power of the two automated models materially increase, the explanatory power of the manual model remains the same. In addition, we find that, whilst our manual measures of negativity and positivity exhibit significant associations with abnormal returns, automated negativity exhibits a significant association, but automated positivity is not significant.

Furthermore, in the light of our findings for the associations of net tone, positivity, and negativity with cumulative abnormal return (CAR) we investigate specific limitations of the automated approach that are the main drivers of this finding. This additional analysis indicates that a) automated measures often double count multiple positive or negative keyword appearing in a single textual clause; b) automated measures cannot detect keywords used in a non-performance context. In both cases, these limitations are more pronounced for positive keyword counts than negative keyword counts.

Additional results indicate that our main findings for net tone still hold when a measure of the length of the IMS or net tone squared is included in the main regression. We also find that the greater explanatory value of manual net tone compared to automated net tone is significant for the smaller firms in the sample, but not for the larger firms in the sample.

Overall, our findings provide the basis for specific recommendations for refining the research designs of large sample studies that are required to rely on automated measures of tone, due to the processing costs associated with scoring tone manually. These specific recommendations are presented in the concluding section.

The remainder of this paper is organized as follows. Section 2 discusses the relative costs and benefits of manual versus automated measures of tone, provides further details on the nature of IMS statements, and justifies the use of abnormal returns around the disclosure of IMSs as a device for comparing manual and automated tone. Section 3 describes the data and sample, explains how we construct our automated and manual tone measures, and presents the main regression models used in the study. Section 4 discusses the main findings of the study, while Section 5 presents some additional results. Section 6 provides concluding remarks and recommendations for improving the use of automated tone measures in future research.

2. Measuring Tone, Interim Management Statements, and the Market Context

2.1. Manual versus Automated Tone

Textual analysis involves analysing the textual or written content of a document, such as the tone of financial disclosures (Henry, 2006, 2008; Loughran & McDonald, 2016). The tone is understood as the sentiment conveyed by the disclosure preparer (i.e. managers), and it signals to users (i.e. investors) whether the disclosure contains fundamentally a positive, neutral or negative message about the firm's current and future economic well-being (Henry, 2008; Henry & Leone, 2016). While early research uses manual analysis techniques to measure tone (e.g. Clatworthy & Jones, 2003; Francis, Philbrick, & Schipper, 1994; Francis, Schipper, & Vincent, 2002; Hoskin, Hughes, & Ricks, 1986; Schleicher & Walker, 2010, 2015), recent research often use computer-assisted word counts (Henry, 2008; Henry & Leone, 2016; Loughran & McDonald, 2011; Neuendorf, 2002). Manual and automated techniques each offer a distinct set of advantages and disadvantages. For instance, while automated analysis is form-oriented and allows quick processing of large volumes of data, manual analysis is context-oriented and able to capture differences in meaning when a word is used in different contexts (Schleicher & Walker, 2010). Consequently, a content analyst's choice between the manual and automated methods is often a trade-off between the perceived accuracy of tone and time convenience.

It has often been suggested in the accounting domain that manual analysis is likely to provide a more accurate measure of the tone than automated analysis because of its perceived ability to capture the contextual meaning (e.g. Clatworthy & Jones, 2003; Schleicher & Walker, 2010). However, empirical evidence of a direct comparison of tone measured by manual and automated methods has not been provided so far. In this study, we address this gap in the literature.

Applying manual textual analysis for tone measurement involves a manual reader determining the sentiment of the written text (e.g. Clatworthy & Jones, 2003; Hoskin et al., 1986; Schleicher, 2012; Schleicher & Walker, 2010, 2015). While the language used in disclosures can vary across firms and industries and over time, managers typically report financial performance in comparative terms (Davis, Piger, & Sedor, 2012) – whether financial performance has improved or deteriorated, or if it is expected to improve or deteriorate. For instance, if it states 'Revenues in the first quarter are likely to be 10% higher than that of last year.', then the content analyst should realize that an *increase* in revenue, a firm fundamental, is likely to positively affect the bottom-line of the firm. The tone of this narrative is therefore ‘Positive’. A disadvantage of manual textual analysis is that it is time consuming and costly to implement. Further, it requires human coders to apply their own judgments, which may be subjective. Therefore, tone measurement quality depends on the human coders’ knowledge and understanding of the business discipline, trends in the industry and their ability in scoring tones objectively and consistently. However, while coder subjectivity can potentially be problematic, given the high inter-coder reliability in prior studies (e.g. Clatworthy & Jones, 2003; Hoskin et al., 1986), manual tone measurement appears to be a relatively simple task (involving basic linguistic ideas such as *good* or *bad, increase* or *decrease*, etc.).

Recent technological advancements have given rise to computer programmes that can provide, for each document processed, the frequency of words from a selected wordlist. Such automated textual analysis treats a document as a ‘bag of words’ (Henry & Leone, 2016). The researcher requires a list of keywords that potentially communicate positive messages, and another list of keywords that potentially communicate negative messages. The software simply returns the number of positive and negative keywords in a document from the two lists, and then the tone is computed, typically, as a scaled difference between the number of positive and negative keywords from the wordlist that appear in the text. An important consideration for measuring the tone is whether all keywords carry equal weights or if some words are weighed more than others. When equal weights are applied to all keywords, if there are more positive than negative words, then the disclosure is said to communicate an overall positive sentiment, i.e. the tone is said to be ‘Positive’ (Henry & Leone, 2016). Although automated analysis can process a large number of documents cheaply and speedily, it is perceived to be less reliable than manual analysis in determining the central meaning of the message (Clatworthy & Jones, 2003; Schleicher & Walker, 2010).

2.2. Interim Management Statements

We use Interim Management Statements (IMSs) for comparing manual against automated tone. IMSs are a policy initiative of the EU that provide trading and financial performance updates for firms listed in the UK and other EU regulated markets for the first and third quarters of the financial year. Introduced via the Transparency Directive (Directive 2004/109/EC), IMSs were adopted by the EU in late 2004 and implemented for UK firms since January 2007. Although IMSs were originally a mandatory disclosure for UK firms, they have been made voluntary since September 2014. The objective of the IMS was to provide for EU firms a simple, low cost alternative to US-style quarterly reports – i.e. a regular disclosure that increases investor confidence and protection with financial performance updates, whilst avoiding the administrative and other disclosure costs associated with full quarterly reports (Schleicher & Walker, 2015). Instead of reporting a quarterly income statement and balance sheet, an IMS meets the Transparency Directive's requirements by providing: (i) a general description of the financial performance and financial position of the firm and (ii) an explanation of the material events and transactions that have taken place during the period. Beyond this, there is no obligation to report specific line-items such as earnings, or to report numbers in the text – the disclosure can be entirely qualitative if the firm so wishes (Link, 2012). In addition, managers have the discretion in determining which events and transactions are 'material'. Therefore, although the IMS was a mandatory disclosure during 2007 – 2014, the content reported in an IMS typically comprise of voluntary textual narratives ('Deloitte & Touché', 2007; Schleicher & Walker, 2015).

As a financial disclosure, the IMS is interesting for several reasons. IMSs are short disclosures, typically one or two pages long, consisting almost entirely of written text. These features make the IMS text corpus an attractive option for full-document manual textual analysis. This contrasts, for example, with lengthy multi-section annual reports which are arguably less suitable for full-document manual analysis, and which often contain graphs, tables, figures and pictures, which are also unsuitable for automated textual analysis. Also, the short textual narrative nature of the IMS makes it an attractive context for comparing manual and automated methods because it enables the tone under each method to be computed based on the same text. Further, IMSs provide financial performance updates, both in relation to very recent trading and (in some cases) looking ahead to the rest of the financial year. Thus, the IMS text corpus is suitable for measuring the tone of financial performance. Finally, there is a growing public interest in IMSs – the Securities and Exchange Commission in the US has recently embarked on a public consultation process about the future of quarterly reporting in the US, and the formal consultation document explicitly refers to the EU IMS experience.

2.3. Market Response as a Benchmark

Prior studies indicate that the tone of financial disclosures such as annual reports (10-K fillings or MD&A sections) and earnings press releases have a strong positive association with contemporaneous abnormal stock returns (Abrahamson & Amir, 1996; Francis et al., 2002; Henry, 2008; Henry & Leone, 2016; Kothari, Li, & Short, 2009; Tetlock, Saar-Tsechansky, & Macskassy, 2008). Therefore, as content analysts weigh up the relative costs and benefits of different approaches to measuring the tone, it is worth comparing their relative explanatory power for abnormal stock returns. As such, Henry and Leone (2016) recently compare the explanatory power for earnings press release tone in the US, measured by different automated wordlists. They find that wordlists specialised for financial communication have greater explanatory power for abnormal stock returns than general non-specialised wordlists. As a follow up, we compare the relative explanatory power of the tone measured by manual and automated analysis for short-window announcement-period abnormal stock returns. This will help content analysts in their assessment of the relative benefits of manual and automated analysis.

However, in order for IMS statements to be a valid context for comparing the alignment of measures of tone with abnormal stock returns, it is important to be sure that such statements exhibit a capacity to convey value relevant information. In this regard, Schleicher and Walker

(2015) report clear evidence of abnormal stock returns on the days surrounding the disclosure of an UK IMS statements. Therefore, it is interesting to consider if the measures of IMS tone are aligned with the abnormal stock returns associated with the publication of the IMS.

Our main analysis involves comparing the relative power of manual and automated net tone models for abnormal stock returns around the release of the IMS. As an additional analysis, we replace the net tone with separate positivity and negativity measures to see if the relative explanatory power of the two methods materially changes. This is because prior studies indicate that the market reaction is more strongly associated with negativity than positivity (Athanasakou, Strong, & Walker, 2017; Tetlock, 2007) and that positive words, but not negative words, are often used in a context that is different from the typical implication of the word (Loughran & McDonald, 2016). Further, Merkl-Davies and Brennan (2007) observe that managers use various impression management and textual subtleties to maximize market rewards for good news and minimize market penalties for bad news. This includes downplaying poor performance and exaggerating good performance, or using tonal words in a way that is not indicative of performance quality. We believe such subtleties are more likely to be revealed by manual than automated analysis.

With respect to our regression analysis, we do not claim causality between tone and market returns. Consistent with prior literature (e.g. Henry & Leone, 2016), we interpret the comparison of the explanatory value of manual and automated tone measures for share price movements as an indicator of the ability of the two approaches to provide a good proxy for the "true" tone of the news in the IMS. However, because the true tone is not observable, we recognise that, in order to be able to draw reliable conclusions about the strengths and limitations of manual and automated tone, we also need to support our market response results by documenting specific differences between the two measures that are likely to drive their associations with true tone.

Therefore, for a carefully chosen subsample of IMSs, we examine the main drivers of the differences between the manual and automated tone by examining the context in which automated keywords are used. Because words may be used differently when the financial performance news is good as opposed to when it is bad, we identify and analyse the IMSs with the largest differences between manual and automated tone when the market returns around the disclosure of IMSs is positive and when the market returns is negative. This analysis reveals the extent to which automated keywords are used in non-financial performance contexts for describing good and bad news.

3. Research Design and Methodology

3.1. Data and Sample Selection

Our sample period spans six years – from 2008 to 2013 when IMSs were mandatory disclosures for firms listed in EU regulated markets. To define a firm-year, we allocate a financial year to the calendar year in which the majority of months falls, and allocate financial years with a June year-end to the calendar year in which the June year-end falls. We use 30 June 2008 as the date for sampling, which had 668 firms in the FTSE All-Share Index. We then eliminate (a) financial firms and (b) firms disclosing full quarterly reports since Article 6 of the EU Transparency Directive indicates such firms do not need to disclose an IMS (Deloitte & Touché, 2007). This leaves 324 non-financial FTSE All-Share Index firms as at 30 June 2008, each requiring the disclosure of two mandatory IMSs every year. From this set of firms, we randomly select 100 firms for textual analysis. This consists of 15 FTSE 100 firms, 38 FTSE 250 firms, and 47 FTSE Small Cap firms, a proportional representation of these indexes from the FTSE All-Share Index population.

We obtain the IMSs from Perfect Information Navigator, a corporate information database that has regulatory and non-regulatory news and filings from over 50,000 firms worldwide. Our sample of 100 firms could potentially yield a maximum of 1,200 IMSs over a six-year period. However, we lose 69 IMSs due to collapse or delisting, and an additional 109 IMSs that were undisclosed by the company, mainly in 2008, the first year of implementation for many listed companies. This leaves a final sample of 1,022 IMSs for manual and automated tone analysis. Table 1 summarizes our sample selection procedure.

[Table 1 near here]

3.2. Tone Measurement

3.2.1. Automated Textual Analysis

For automated textual analysis we use a publicly available computer software tool named the *Corporate Financial Information Environment – Final Report Structure Extractor* (CFIE-FRSE). This software tool is the outcome of a publicly funded collaboration between Lancaster University Management School, Alliance Manchester Business School, and the London School of Economics, and it is available at http://ucrel.lancs.ac.uk/cfie/cfie-frse-software.php. The software tool processes the textual content of documents in either PDF, HTML or plain text format. A list of keywords is first uploaded into CFIE-FRSE, in plain text format. Then the document with the textual content is uploaded. Based on the wordlist, the CFIE-FRSE tool screens the textual corpus of the document and returns the total frequency of words in the document that matches with the words in the wordlist. The retrieved text of the document is processed automatically after uploading and outputs appear in an Excel spreadsheet. The CFIE-FRSE tool also returns the total document word count by default (El-Haj, Alves, Rayson, Walker & Young, 2018).[2]

[2] The CFIE-FRSE tool also returns in the Excel spreadsheet the following information by default: (i) total page count (for multiple document files) (ii) Fog index of readability (iii) Flesch-Kincaid index of readability (iv) counts for positive and negative words of the Loughran and McDonald (2011) wordlist (v) count of forward-looking words (vi) count of strategy related words (vii) count of uncertainty-related words and (viii) count of causal-reasoning words. For further details on the mechanism of CFIE-FRSE please see El-Haj et al. (2018).

We use the two best known automated wordlists specialized for financial and business communication: (i) Henry (2008) and (ii) Loughran and McDonald (2011). The Henry (2008) wordlist includes a total of 105 positive words and 85 negative words.[3] The Loughran and McDonald (2011) wordlist contains 354 positive words and 2355 negative words. We upload all IMSs in CFIE-FRSE and obtain the number of positive and negative words from the Henry (2008) and Loughran and McDonald (2011) wordlists present in each IMS document. Since this is a speedy procedure, the word counts in each list for all 1,022 IMSs were obtained within a day. Then, separately for each wordlist, we calculate the automated net tone score for every IMS:

$$\mathrm{TONE}_W = (\mathrm{POSITIVE}_W - \mathrm{NEGATIVE}_W) / (\mathrm{POSITIVE}_W + \mathrm{NEGATIVE}_W) \qquad (1)$$

In the above formula, $\mathrm{POSITIVE}_W$ and $\mathrm{NEGATIVE}_W$ refer to the word frequency from CFIE-FRSE based on the positive and negative words in the automated wordlists with w = {Henry, LM}.[4]

The automated net tone scores of the two lists, TONE_{HENRY} and TONE_{LM}, are continuous variables ranging from totally negative (-1) to totally positive (1). For any of these lists, if there are more negative (positive) words than positive (negative) words in the IMS document, then the corresponding net tone score would range between -1 and 0 (0 and 1), and would indicate that the IMS depicts an negative (positive) sentiment. The absence of any negative (positive) words would make the tone 1 (-1). A tone score of zero can be achieved if the number of positive or negative words in the IMS is equal.

[3] Henry and Leone (2016) use a wordlist that has 188 positive words and 93 negative words. However, we use the Henry (2008) wordlist because: a) it has become the standard automated wordlist for financial domain; b) it is easily available; c) most published studies using Henry's wordlists have used the 2008 wordlist.

[4] We employ equal weighting for the positive and negative words in the automated wordlists. Henry and Leone (2016) prescribe equal weighting because it is simple, intuitive, easy to implement, and find that inverse document frequency (*idf*) weighting provides no improvement over equal weighting.

For each wordlist, we also compute separate positivity and negativity scores. Positivity (negativity) is the number of positive (negative) words from the wordlists divided by the total number of words in the entire IMS document. We term the positivity measures POS_{HENRY} and POS_{LM} and the negativity measures NEG_{HENRY} and NEG_{LM} denoting the Henry and LM wordlists.[5] Note that these positivity and negativity measures all range from 0 to 1.

3.2.2. Manual Textual Analysis

Our unit of manual textual analysis is the clause, not the sentence. For the purpose of manual analysis, we define a clause as a phrase or group of words that contains distinct information on a specific topic. Thus, clauses typically contain a subject (topic) and a predicate (description of action). While most clauses in our analysis are complete single textual sentences, occasionally: (i) a sentence contains more than one clause (if multiple topics or more than one piece of information is discussed within one textual sentence) (ii) a clause comprises of multiple sentences (in the case of repeated sentences, such as statements highlighted early on in bullet points but also repeated in the text subsequently).

We conduct manual textual analysis on the selected sample by reading each IMS and recording on an Excel spreadsheet, the tone of every single clause in the IMS – whether it is 'Positive', 'Neutral' or 'Negative'. Therefore, the text corpus is the same as that processed by CFIE-FRSE for the automated wordlists, enhancing comparability of the methods. One of the authors performed the manual analysis, after a series of iteration rounds with small samples of IMSs coded by the author and checked by the other authors to see if they agreed with the coding and tone assignment. Disagreements were mutually discussed and resolved. As a rule, 'Positive' clauses are those that have clear or direct indications of improvement or progress from the previous circumstance (e.g. 'Profit in first quarter this year was 10% *higher* than the

[5] Following Loughran and McDonald (2011), for automated scoring, if there is a negation of words immediately before a positive word, we count the positive word as negative. The words that we use for identifying negated positives are 'no', 'not', 'none', 'neither', 'never', and 'nobody'. We find that the incidence of negated positive words is less than 1% of all negative words and the results are qualitatively similar whether these words are counted as negatives or positives.

corresponding quarter last year'). These clauses typically convey good news. 'Negative' clauses are those that provide clear or direct indications of 'deterioration' from the previous circumstance, (e.g. 'Group revenue for the year is expected to further *decrease* as challenges in the economy have not subsided'). These clauses might be called bad news. 'Neutral' clauses are those that exhibit the following characteristics: (i) they are neither distinctly positive nor negative, (ii) performance is in line with expectations, (iii) the status quo is preserved (e.g. 'Our trading performance in the third quarter has been satisfactory, *in line* with our expectations announced during the half-yearly results') or (iv) any clause which are not directly related to the firm's financial or economic well-being. It takes about 850 hours of manual coding to return the number of positive and negative clauses for all 1,022 IMSs, including the pilot studies and iteration rounds.[6] In other words, on average, it takes about 50 minutes to score one IMS.

After determining the manual tone for each clause, we compute the overall tone score for the entire IMS document. In the scoring spreadsheet, each clause with a Positive (Negative, Neutral) tone is coded as an indicator variable which takes the value of 1 if the clause is positive (negative, neutral), and zero otherwise. We then compute the manual positivity (negativity) of an IMS, $\text{POS}_{\mathit{MANUAL}}$ ($\text{NEG}_{\mathit{MANUAL}}$), by dividing the total number of positive (negative) clauses in the IMS with the sum of all positive, negative and neutral clauses in the IMS. The manual positivity (negativity) computed in the above manner indicates the proportion of positive (negative) clauses out of all the clauses in an IMS. $\text{POS}_{\mathit{MANUAL}}$ and $\text{NEG}_{\mathit{MANUAL}}$ can range from 0 to 1 and the inclusion of neutral clauses in the denominator makes the manual positivity and negativity measures like their automated counterparts and also eliminates any potential problems of linear dependency if these measures are used together in a regression model. Finally, we compute the manual net tone score, $\text{TONE}_{\mathit{MANUAL}}$, for each IMS as the difference

[6] 850 hours of manual coding loosely translates to eight hours of coding for five days a week, over a period of five months.

between the number of positive and negative clauses in an IMS, $\text{POSITIVE}_{\text{S}}$ and $\text{NEGATIVE}_{\text{S}}$ respectively, divided by the sum of positive and negative clauses in the entire IMS, as follows:[7]

$$\text{TONE}_{\textit{MANUAL}} = (\text{POSITIVE}_{\text{C}} - \text{NEGATIVE}_{\text{C}}) / (\text{POSITIVE}_{\text{C}} + \text{NEGATIVE}_{\text{C}}) \qquad (2)$$

$\text{TONE}_{\textit{MANUAL}}$ is a continuous variable that ranges from totally negative (-1) to totally positive (1). The automated tone measures $\text{TONE}_{\textit{HENRY}}$ and $\text{TONE}_{\textit{LM}}$ are similar to the manual tone measure $\text{TONE}_{\textit{MANUAL}}$ in that they all have the same range of possible values (-1, 1). If there are more negative (positive) clauses than positive (negative) clauses in an IMS, then $\text{TONE}_{\textit{MANUAL}}$ would range between -1 and 0 (0 and 1), indicating that the overall sentiment in IMS is negative (positive). The absence of any negative (positive) clauses would make the IMS tone 1 (-1). A net tone score of zero is recorded if the number of positive clauses in the IMS is equal to the number of negative clauses. Appendix A provides some examples of the differences between manual and automated scores. Appendix B presents some high frequency positive and negative words from the Henry and LM wordlists and compares them to the manual tone of the clauses they present.

3.3. Regression Models

As we cannot observe the true underlying tone that reflects the managers' sentiment about the financial performance, following Henry and Leone (2016), we assume semi-strong form market efficiency and compare the extent to which the alternative tone measures are aligned most closely with the abnormal stock returns around the release of the IMS. For this, we follow prior literature (e.g. Davis et al., 2012; Henry, 2008; Henry & Leone, 2016) and employ a short-window announcement-period event study methodology. Short-window event

[7] For tone computation, we assign equal weights to every clause similar to the automated methods. However, we aggregate the tone at the word level for $\text{TONE}_{\textit{HENRY}}$ and $\text{TONE}_{\textit{LM}}$ but at the clause level for $\text{TONE}_{\textit{MANUAL}}$ to ensure that they have the same range of values. Our aggregation is consistent with the respective literature of manual and automated scoring (e.g. Clatworthy & Jones, 2003; Henry & Leone, 2016; Schleicher & Walker, 2010).

studies are widely accepted in capital markets research for their reliability in measuring the market response to the release of financial disclosures (Lougee & Marquardt, 2004; Schrand & Walther, 2000) and provide evidence for understanding corporate policy decisions, such as the association between tone and abnormal stock returns (Henry, 2008). Further, given that semi-strong form market efficiency implies that share prices adjust quickly to the disclosure of all new public information (Kothari & Warner, 2006) a short-window around the IMS disclosure date is likely to capture most of the relevant market reaction. In our main model we regress CAR on alternative measures of tone as follows (omitting year and industry fixed effects):

$$CAR = \alpha + \beta_1 TONE + \beta_2 SIZE + \beta_3 BTM + \beta_4 LOSS + \varepsilon \quad (3)$$

For supplementary analysis, we then replace the overall tone, TONE, with separate positivity and negativity measures, POS and NEG, as follows:

$$CAR = \alpha + \beta_1 POS + \beta_2 NEG + \beta_3 SIZE + \beta_4 BTM + \beta_5 LOSS + \varepsilon \quad (4)$$

In the above models, the short-window abnormal stock return around the IMS announcement date is measured by the three-day cumulative abnormal return, CAR. For abnormal returns, we calculate daily market model adjusted returns, u_{it}, as $u_{it} = R_{it} – (\alpha_i + \beta_i R_{mt})$, where R_{it} is the return of firm *i* on day *t*, R_{mt} is the return of the FTSE All-Share Index on day *t* and where R_{it} and R_{mt} are calculated from DataStream Return Indices, RI. α_i and β_i are firm *i*'s estimated market model parameters calculated from the non-event period which runs from day t–60 to day t–10 and from day t+10 to day t+60 relative to the IMS announcement day t =

0. CAR is calculated as the sum of the daily market model adjusted returns, u_{it}, over the three-day event period (days t–1, t, t+1), such that $CAR_{it} = u_{it-1} + u_{it} + u_{it+1}$.[8]

A few recent studies have replaced the net tone with the periodic change in tone in their modelling of abnormal stock returns (e.g. Davis et al., 2012; Henry & Leone, 2016). In Equation 3 (and throughout the paper) we follow the bulk of the tone literature and use levels in net tone, not change in net tone (e.g. Henry, 2008; Tetlock et al., 2008). This is because we believe the net tone itself is an incremental figure and not an accumulation of past sentiment; computing periodic changes in tone may lead to double computation of the *change* in sentiment. For instance, we observe that the majority of words in the Henry and LM wordlists are either verbs or adverbs while the remaining are adjectives. Verbs and adverbs either indicate a change from prior circumstance or characterize a sentiment, hence they do not accumulate tone from prior disclosures, but rather represent incremental change in sentiment (e.g. *increase, decline, grew, adversely, lower*). Additionally, adjectives that are accompanied by time or performance benchmarks such as managerial expectations, analyst or market consensus, prior periods, full-year guidance, etc. also indicate changes in sentiment (e.g. ‘We experienced *robust* sales performance in the quarter, relative to our prior expectations’). We further argue that adjectives without explicit benchmarks may also imply a change in sentiment. For instance, the clause ‘Trading during the period remained *poor*’ uses the adjective *poor* without an explicit benchmark, but is arguably used here relative to some implied standard—if ‘trading’ is indeed poor, it must be poor relative to some benchmark. Therefore, we believe the net tone, not periodic changes in tone, should be used as a regressor in market return estimations.

[8] We believe the three-day CAR around the IMS disclosure date (days t–1, t, t+1) is an appropriate measure of short-window announcement-period share price reaction because plotting the mean and median daily absolute abnormal return for eleven days surrounding the IMS disclosure date (days t–5 ... t+5) indicates that peaks of both mean and median absolute abnormal return occur on the IMS disclosure date, with no signs of elevated reaction on the surrounding days. This is suggestive of an instantaneous market reaction with no prior leakage of information.

To address the difference in the distribution of the three tone measures, we follow Henry & Leone (2016) and standardize all tone, positivity and negativity variables to have a mean of 0 and standard deviation of 1. We use these standardized tone values in our regression estimations to allow the tone coefficients to be compared directly to one another across different models, though it is important to note that we compare manual versus automated models primarily on the basis of explanatory power, that is, adjusted R-Squared.

Guided by prior literature (e.g. Henry, 2008; Henry & Leone, 2016) we include several control variables in Equations (3) and (4): i) firm size, (SIZE), the natural logarithm of the market value of equity at the beginning of the year *t*, calculated as the number of shares multiplied by share price, both at the start of the year t; ii) book-to-market value, (BTM), calculated as the ratio of the book value of equity to the market value of equity at the start of the year; iii) profitability status, (LOSS), an indicator variable which equals 1 if pre-exceptional operating profit is less than zero at the start of the year, and 0 otherwise. Also included in each of the models are five indicator variables for the six years in the sample period, omitting 2008, and eight indicator variables for the nine ICB industry classifications for FTSE All-Share firms, omitting ‘Industrials’. All variables are defined in Table 2.

[Table 2 near here]

A relevant research design question is the potential inclusion of unexpected earnings, typically computed as the difference between actual and forecasted EPS and scaled by share price, to control for information in reported net income. However, unlike Davis et al. (2012) and Henry & Leone (2016) who examine earnings press releases, we do not control for unexpected earnings in our models since IMSs are not earnings announcements and do not contain income statements. The IMS content is largely qualitative written text, and therefore, we believe it is not necessary to control for news outside the narratives in an IMS, as it risks distorting the tone coefficients.[9]

It is also important to stress that the focus of our study is on the comparative alignment of manual and automated measures of tone with short-term price movements. Thus, we do not claim that our research design is capable of demonstrating a causal link from tone to abnormal stock returns. Any attempt to do so would require a research design that is capable of addressing the possibility of endogeneity due to correlated missing variables and measurement error in the dependent variables and measures of tone. This is beyond the scope of the present study.[10]

4. Results

4.1. Descriptive Statistics

Table 3 presents the descriptive statistics for all 1,022 IMS observations. Statistics for tone and positivity and negativity measures are shown prior to standardization. The means of all three net tone measures are positive. This is consistent with Rutherford's (2005) assertion that financial narrative disclosures typically contain a greater proportion of positive messages when managers have discretion over the content, as in the case of IMSs. All three net tone scores have a maximum of 1 and minimum of –1.

The mean and median net tone scores for both of the automated measures ($\text{TONE}_{\textit{HENRY}}$ mean=0.59 median=0.64; $\text{TONE}_{\textit{LM}}$ mean=0.51 median=0.55) are higher than those for manual tone ($\text{TONE}_{\textit{MANUAL}}$ mean=0.25 median=0.27). Comparing each automated list with the manual scores, we find, in both cases, that a t-test for the difference in means yields a p-value of 0.000

[9] For instance, Schleicher & Walker (2015) find that only 4% of forward-looking earnings and 20% of backward-looking earnings information reported in IMSs is quantitative in nature, and is almost always embedded within the narrative.

[10] In the context of a study of the comparative alignment of manual versus automated tone with CAR, endogeneity would be a concern only if there is a good reason to believe that it could have a differential effect on automated versus manual measures of tone. We can think of no reason why this should be the case. In particular, it is most unlikely that errors in scoring tone by either manual or automated methods, will be correlated with economic factors driving the market response to the IMS.

as does a Wilcoxon rank-sum test for the difference in medians. The significant differences between manual and automated tone scores is consistent with the findings of Abrahamson and Amir (1996) that narrative disclosures often include an excessive number of positive words, not all of which may convey a meaning of improved financial performance.

The mean and median of both positivity and negativity in the LM wordlist are about two times larger than the Henry wordlist, perhaps due to the greater number of positive and negative words in the LM wordlist than in the Henry wordlist. In both automated lists, the mean value of positivity is greater than that of negativity. This is also consistent with prior literature (Abrahamson & Amir, 1996; Tetlock, 2007).[11] Additionally, the descriptive statistics for manual indicate that there are more positive clauses than negative clauses in an IMS. The average length of an IMS is 1,010 words or 28 clauses, as opposed to the median length of 777 words or 22 clauses, reflecting a degree of right skewness in the length of the IMSs in the sample.[12]

[Table 3 near here]

4.2. Comparative Distributions of the Tone Measures and CAR

Table 4 provides detailed comparisons of the distributions of the three tone measures and CAR. Panel A presents selected percentiles of the four distributions and the number of observations greater than or less than zero. We observe that CAR is evenly distributed around zero with half the values greater than zero and half the values less than or equal to zero. The

[11] We obtain a slightly higher mean and median net tone score than Henry & Leone (2009, 2016) due to a slightly greater number of positive words and marginally lower number of negative words per document in our sample as opposed to theirs. This might be attributed to differences in disclosure regulation and the culture of litigation, that is, a greater number of lawsuits are filed in US as opposed to UK causing US firms to be less optimistic and more cautious in their wording as opposed to UK firms.

[12] We observe from manual analysis that these include manufacturing firms that disclose their quarterly production results in IMSs, consistent with the objective of an IMS to inform investors about updating material events and transactions as part of reporting financial performance.

percentiles of CAR indicate that the distribution of the positive values is roughly similar to that of the negative values. All three tone distributions have median values that are materially greater than zero. This is especially the case for the two automated tone measures. For all three tone measures, the number of observations greater than zero is much greater than the number of negative observations. In particular, less than 5% of the automated tone observations are negative consistent with a strong bias towards positive tone.

Panel B reports the deciles of the three tone measures. For each decile, we take the observations corresponding to that decile, and then calculate the median value of the tone measure for which the decile is being reported, and the median values of the other tone measures and CAR for the same set of observations. The results indicate that, across the manual tone deciles, the median values of the three tone measures increase monotonically with each other. Therefore, the three tone measures give roughly similar orderings of tone in terms of decile medians. We also see a rough correspondence between the tone deciles and the median CAR values. For $\text{TONE}_{\mathit{MANUAL}}$, the highest (lowest) decile corresponds to the highest (next to lowest) CAR. Also, the three highest (lowest) manual tone deciles correspond to the three highest (lowest) CARs. For $\text{TONE}_{\mathit{HENRY}}$, the second highest (lowest) decile corresponds to the highest (lowest) CAR. The lowest five deciles exhibit a monotonic relation with CAR, but a monotonic relation between decile and CAR breaks down for the five highest deciles. For $\text{TONE}_{\mathit{LM}}$, the second lowest (third highest) decile corresponds to the lowest (highest) CAR. The relation with CAR for both the five lowest deciles and the five highest deciles are not monotonic.

[Table 4 near here]

4.3. Correlation Table

Table 5 presents Spearman's rank correlations between the tone, positivity and negativity measures, and the other variables used in our study. CAR, the measure of abnormal stock returns, has a stronger positive correlation with TONE$_{MANUAL}$ (r=0.21) than with either TONE$_{HENRY}$ (r=0.12) or TONE$_{LM}$ (r=0.15). This provides some prima facie evidence that manual tone is more closely associated with abnormal market reactions than automated tone, something we noted already from Table 4, Panel B. CAR is also positively (negatively) associated with the manual and automated positivity (negativity) measures, and, while these correlations are all significant at the 1% level, the correlations are stronger for the manual positivity and negativity measures, suggesting that manual textual analysis is better at picking up the good and bad news messages conveyed in an IMS. The absolute magnitudes of CAR's correlations with the negativity measures are greater than the correlations with the positivity measures for manual and both automated lists, consistent with Tetlock (2007). This implies that the instantaneous market response to an IMS disclosure is more strongly associated with negativity than positivity. CAR is more strongly and negatively (positively) associated with the manual negativity (positivity) measures than with either of the automated wordlists. This provides preliminary evidence that manual positivity and negativity are more strongly associated with market returns than automated positivity and negativity.

The intra-tone correlations all exhibit expected associations. For instance, TONE$_{MANUAL}$ is positively associated with TONE$_{HENRY}$ (r=0.58) and TONE$_{LM}$ (r=0.61) but the two automated measures have much stronger correlation (r=0.89). TONE$_{MANUAL}$ is positively (negatively) associated with all three measures of positivity (negativity), but the correlations are greater for the manual measures of positivity (negativity) than the corresponding automated measures. The same applies for the automated measures – both TONE$_{HENRY}$ and TONE$_{LM}$ have stronger associations with their own positivity and negativity measures than with other measures. Finally, the automated measures of positivity and negativity have stronger correlations with each other than with their manual counterparts.

[Table 5 near here]

4.4. The Associations of CAR with Manual and Automated Tone

Table 6 presents the regressions of CAR on the manual and automated tone measures for our full sample of 1,022 IMSs.[13] In Panel A, we observe that TONE$_{LM}$ and TONE$_{MANUAL}$ are positively associated with the CAR, with both being significant at the 5% level. The positive alignment between tone and market returns is consistent with prior literature (e.g. Henry, 2008; Henry & Leone, 2016; Loughran & McDonald, 2011). Additionally, we find that the manual tone model has a larger adjusted R-Squared than either of the automated models.[14] The Vuong (1989) tests in Panel B confirm that manual tone has greater explanatory power for abnormal stock returns than either Henry tone (p=0.025) or LM tone (p=0.032).

It is possible that our results could vary with the components of the net tone. In particular, there could be differences in the explanatory power of the positive and negative tone components of tone. However, since manual analysis can capture context, we expect both the manual positivity and negativity to be closely aligned with CAR. Therefore, replacing the net tone with separate positivity and negativity measures is unlikely to materially increase the explanatory power of manual tone models. In contrast, prior research suggests that while positive words are frequently used in non-positive contexts, negative words are less frequently used in non-negative contexts (Loughran & McDonald, 2011). As such, we would expect automated negativity, but not positivity, to be strongly aligned to CAR. When the net tone is replaced by separate positivity and negativity measures, it is possible that the inclusion of the

[13] The reported results control for industry and year fixed-effects. Replacing year fixed-effects with quarter fixed-effects yields qualitatively similar results.

[14] We observe that the R-Squared values for CAR is quite low, particularly in the automated net tone models, consistent with the suggestion that in terms of discretionary content the typical IMS is similar to a trading statement, not an earnings announcement, focusing more on sales and trading than the bottom-line.

more context-accurate negativity measure increases the explanatory power of the automated methods.

To examine these possibilities, we replace the net tone with positivity and negativity in all three models. In Panel A, we find that manual negativity (positivity) is negatively (positively) associated with CAR at the 5% significance level. However, for both automated models, whilst negativity is significantly associated with CAR at the 5% level, positivity has no significant association. The adjusted R-Squared values for both of the automated models improves materially with the inclusion of separate positivity and negativity, but remains largely unchanged in the manual model, consistent with our expectations. The Vuong (1989) tests in Panel B indicate that manual tone model now has marginally greater explanatory power than the Henry tone model (p=0.099) but no significantly greater explanatory value than the LM tone model (p=0.234).

[Table 6 near here]

A related question involves examining whether the tone of positive (negative) automated word counts exhibit the same explanatory power for CAR as the tone of manually scored positive (negative) clauses. The results are un-tabulated for brevity. In CAR regressions of positivity-only models, we find, while manual positivity is positively associated with CAR, neither of the automated positivity measures have a significant association. In contrast, in negativity-only models, both manual and automated negativity are negatively associated with CAR, consistent with Table 6. Vuong (1989) tests of positivity-only and negativity-only models indicate that the manual positivity-only model has greater explanatory power than its positivity-only automated counterparts, but the differences are insignificant for the negativity-only measures.

Finally, for all three tone measures, we examine the change in model explanatory power when the net tone is replaced with positivity and negativity. The results are again un-tabulated for brevity. Vuong (1989) tests indicate that for both the Henry and LM wordlists, the positivity and negativity model has significantly greater explanatory power than the net tone model (p=0.000 and p=0.000). However, there is no statistically significant difference in the explanatory power of the two manual models (p=0.810).

Overall, the results in Table 6, together with the subsequent untabulated results, indicate that manual measures of net tone clearly provide a better basis for explaining abnormal returns than automated measures of net tone. However, when the net tone is replaced by positivity and negativity, we find that the superiority of manual tone is significant for positivity but not for negativity. This latter finding is likely due to higher degrees of bias and noise in the automated measures with respect to the scoring of positive news. Crucially, when positivity and negativity are put together in the same model, the differences in explanatory power between manual and automated methods largely disappear. Further analysis of the deficiencies of the automated measures is presented in the next section.

5. Additional Analyses

In this section, we report the results of additional tests that provide further insights into the main differences between automated and manual tone measures, explore the sensitivity of the main results to firm size and introduce additional controls into our models for the IMS length and tone squared.

5.1. Key Drivers of the Difference between Manual and Automated Tone

We analyse the drivers of the most material differences between manual and automatic tone measures by examining (i) the extent to which multiple automated keywords are used in a single clause and (ii) the context in which the tonal words are used. As an aid to understanding the most material differences between manual and automated tone, we identify the IMSs that have the largest differences between manual and automated tone when the market returns are positive and when the market returns are negative, by using the following pair of regressions:

$$\text{TONE}_{\mathit{MANUAL}} = \alpha + \beta_1\text{TONE}_{\mathit{HENRY}} + \beta_2\text{CARNEG} + \beta_3 (\text{CARNEG} \times \text{TONE}_{\mathit{HENRY}}) + \varepsilon \qquad (5a)$$

$$\text{TONE}_{\mathit{MANUAL}} = \alpha + \beta_1\text{TONE}_{\mathit{LM}} + \beta_2\text{CARNEG} + \beta_3 (\text{CARNEG} \times \text{TONE}_{\mathit{LM}}) + \varepsilon \qquad (5b)$$

In the above regressions, CARNEG = 1 if CAR less than or equal to zero. The residual, ε, in each of the above two regressions identifies the magnitude of the difference between manual and automated tone when CAR is positive and when CAR is negative. Then, for each regression, we select for analysis the IMSs with the 10 largest positive and the 10 largest negative residuals. In other words, we select the IMSs with the largest deviations (residuals) between manual and automated tone, controlling for CAR. For each automated tone measure, *HENRY* and *LM*, this procedure selects 40 IMSs in total for detailed analysis and comparison that is, ten for relatively high manual (HIGHMAN) when CAR is negative (CARNEG), ten for relatively high manual (HIGHMAN) when CAR is positive (CARPOS), ten for relatively low manual (LOWMAN) when CAR is negative (CARNEG), and ten for relatively low manual (LOWMAN) when CAR is positive (CARPOS).

The results are reported in Table 7 where the findings of our analysis are shown separately for HENRY and LM tone. Column (1) shows the total number of clauses in the 10 IMS statements analysed. Column (2) shows the number of clauses containing none of the automated tone keywords. Column (3) shows the number of clauses containing one or more of the automated tone keywords. Columns (4) & (5) show the number of clauses containing two or more positive and two or more negative keywords, respectively. Columns (6) & (7) show the number of clauses where at least one positive and at least one negative keyword, respectively, is used in a context that is unrelated to firm financial performance (and hence where manual analysis assigns a neutral tone).

The overall results, summarised in the last two rows of the Table 7, indicate that multiple positives occur much more frequently than multiple negatives, both absolutely and relative to the number of clauses in which any of the keywords appear. In addition, in columns (6) and (7) we see that, for both positive and negative keywords, there are a material number of cases where keywords are used in a non-performance related context. We see that the frequency of the use of positive keywords out of context is greater than the frequency of negative keywords out of context, both absolutely and relative to column (3).

The higher relative frequency of multiple positive keywords, compared to multiple negatives, holds for both of the TOTAL – HIGHMAN rows and the TOTAL – LOWMAN rows. However, there is a material difference in the relative frequency in the use of negative to positive keywords out of context between TOTAL – HIGHMAN and TOTAL – LOWMAN. For TOTAL – HIGHMAN (TOTAL – LOWMAN) the absolute and relative use of negative (positive) keywords out of context is greater than the use of positive (negative) keywords out of context.

In Table 7, it is also of some interest to compare the TOTAL results for HENRY against LM. The number of clauses in columns (4) to (7) is materially higher for LM than HENRY, almost certainly due to the LM word lists being much longer than the HENRY word lists. However, when we compare the percentages for the TOTAL HENRY and LM results, we see that the relative use of multiple positives is slightly higher for HENRY than for LM. For multiple negatives, the LM percentage is almost double the HENRY percentages, perhaps due to the large number of negative keywords in the LM list. Both of the percentages for keywords used out of context are materially higher for LM compared to HENRY.

[Table 7 near here]

5.2. Partitioning the Sample by Firm Size

We investigate if the alignment between tone and market returns is different for the small firms in the sample. Nearly half (47%) of the sample firms are in the FTSE Small Cap index. Smaller firms are likely to have lower visibility and lower analyst following (Bhushan, 1989) leading to lower stock liquidity (Cheung & Ng, 1992). Consequently, it is possible that the IMS tone has a stronger alignment with CAR for small firms than for large firms, as lower analyst research activity might imply less pre-IMS anticipation of IMS trading results, and hence more surprise and more market reaction on IMS announcement dates. To examine this prediction we group our IMS observations into two subsamples – large firms, that is, IMSs disclosed by FTSE 100 or FTSE 250 firms, and small firms, that is, IMSs disclosed by FTSE Small Cap firms, and we regress CAR on tone separately for each of these two groups.

The results are reported in Table 8. In Panel A, we find, consistent with our expectation, that the coefficients on all three net tone measures are higher (lower) for small firms (large firms) when compared against the respective full sample coefficient in Table 6. In Panel B, Vuong (1989) tests indicate that, for small firms, the manual tone model has greater explanatory power for abnormal stock returns than both the Henry tone model (p=0.008) and the LM tone model (p=0.021). In contrast, for large firms, the manual tone model has marginally greater explanatory power than the Henry model (p=0.081) but not the LM model (p=0.724). Overall, these results indicate that the greater explanatory power of the manual tone models relative to automated tone models is much stronger for small firms than for large firms.

[Table 8 near here]

5.3. Controlling for Length and Tone Squared

Following Henry (2008), we examine the sensitivity of our findings for net tone to the inclusion of, respectively, the length of the IMS statement and the squared value of net tone in the main model. First, we add the IMS length, LENGTH (defined as the natural logarithm of the number of words in the IMS document) as an additional explanatory variable to our main models. The results are un-tabulated for brevity. We find that LENGTH is negatively aligned with CAR in all three models, significant at the 1% level for the Henry model and at the 5% level for the LM and manual tone models. $\text{TONE}_{\mathit{HENRY}}$ is positive and insignificant, $\text{TONE}_{\mathit{LM}}$ is positive and significant at the 5% level, and $\text{TONE}_{\mathit{MANUAL}}$ is positive and significant at the 1% percent level. Vuong (1989) tests indicate that the $\text{TONE}_{\mathit{MANUAL}}$ model continues to have significantly greater explanatory power than either the Henry tone model or the LM tone model.

Second, Henry (2008) suggests that the relationship between tone and market returns is non-linear. To see if the same holds true for our models, we add as an explanatory variable TONESQUARED (defined as the squared value of net tone) to our models. In un-tabulated results, we find that TONESQUARED is positive and significant at the 1% level in the Henry tone model (in contrast to the negative and significant finding in Henry (2008)) but insignificant for the other two tone measures. Both $\text{TONE}_{\mathit{HENRY}}$ and $\text{TONE}_{\mathit{MANUAL}}$ are now positive and significant at the 1% level, while $\text{TONE}_{\mathit{LM}}$ is no longer significant. Vuong (1989) tests indicate that the manual tone model continues to have greater explanatory power than either of the automated models.

6. Concluding Remarks

We contribute to the debate on manual versus automated methods of measuring tone by providing evidence that manual net tone has greater alignment with abnormal stock returns around the disclosure of UK IMSs. An additional analysis indicates that the alignment of automated tone measures, but not of manual tone, improves when measures of the net tone are replaced by separate measures of positivity and negativity. Indeed, the differences in explanatory power between the manual and automated methods disappear when separate measures of positivity and negativity are used instead of the net tone. However, we also find that automated measures of positivity exhibit no significant association with market movements, consistent with such measure containing a considerable amount of noise.

Detailed comparisons of the net tone measures indicates that while all three measures are positively biased, the two automated measures are materially more biased than the manual measure. Additionally, an analysis of 80 IMSs which display the largest differences between manual and automated tone indicates that manual tone differs from automated tone because of the use of multiple automated keywords in one clause and the use of keywords in a non-performance context.

Overall, our results suggest that, in research contexts where it is too costly to use manual measures of tone for the full sample, researchers may be able to improve the use of automated word lists by (i) using separate measures of positivity and negativity in their regression models; (ii) tailoring the standard word lists to their specific research context by removing from the positive and negative word lists the keywords that most frequently occur in clauses containing more than one keyword; and (iii) controlling for document length and tone squared in their models. Furthermore, if time is available to calculate tone scores manually for a subsample of documents, then it may be possible to use this analysis to identify keywords that are frequently used out of the context of the study with a view to excluding them from the standard automated lists. In most research contexts, it will make sense to test if the results differ across firm size groups, and to document the sensitivity of the research findings to controlling for document length.

There are a number of avenues for future research. While this paper has focused on comparing manual versus automated measures of the tone, there are other linguistic features, such as attribution analysis and textual readability for which analogous comparisons can be made. We also believe that the tonal analysis of other price sensitive company documents could be of interest, such as earnings announcements and conference call transcripts. Our results for the specific case of IMSs suggest that it may be possible to tailor automated methods to specific corpus of narratives by using insights gained from the manual analysis of a preliminary subsample of the corpus.

References

Abrahamson, E., & Amir, E. (1996). The information content of the president’s letter to shareholders. *Journal of Business, Finance and Accounting, 23*(8), 1157–1182.

Athanasakou, V., Strong, N.C., & Walker, M. (2017). *Asymmetric Information Flows*. Working Paper: London School of Economics.

Bhushan, R. (1989). Firm characteristics and analyst following. *Journal of Accounting and Economics, 11*(2-3), 255–274.

Cheung, Y., & Ng, L.K. (1992). Stock prize dynamics and firm size: An empirical investigation. *The Journal of Finance, 47*(5), 1985–1997.

Clatworthy, M., & Jones, M.J. (2003). Financial reporting of good news and bad news: Evidence from accounting narratives. *Accounting and Business Research, 33*(3), 171–185.

Davis, A.K., Piger, J.M., & Sedor, L.M. (2012). Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research, 29*(3), 845–868.

Deloitte & Touché. (2007). First IMpressionS: The First Year’s Interim Management Statements. London: The Creative Studio at Deloitte.

El-Haj, M., Alves, P., Rayson, P., Walker, M., & Young, S. (2018). *Retrieving, Classifying and Analysing Narrative Commentary in Unclassified (Glossy) Annual Reports.* Working Paper: Lancaster University.

Francis, J., Philbrick, D., & Schipper, K. (1994). Shareholder litigation and corporate disclosures. *Journal of Accounting Research, 32*(2), 137–164.

Francis, J., Schipper, K. & Vincent, L. (2002). Expanded disclosures and increased usefulness of earnings announcements. *The Accounting Review, 30*(1), 185–209.

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), 1–19.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication, 45*(4), 363–407.

Henry, E., & Leone, A.J. (2009). Early working version of Henry and Leone (2016) titled ‘Measuring qualitative information in capital markets research’. available on the internet at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1470807 accessed 15.08.2014.

Henry, E., & Leone, A.J. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review, 91*(1), 153–178.

Hoskin, R.E., Hughes, J.S., & Ricks, W.E. (1986). Evidence on the incremental information content of additional firm disclosures made concurrently with earnings. *Journal of Accounting Research, 24*(1), 1-32.

Kothari, S.P., Li, X., & Short, J. (2009). The effect of disclosures by management, analysts, and financial press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review, 84*(5), 1639–1670.

Kothari, S.P., & Warner, J.B. (2006). Econometrics of event studies. *Working paper, Center for Corporate Governance, Tuck School of Business at Dartmouth.*

Link, B. (2012). The struggle for a common interim frequency regime in Europe. *Accounting in Europe, 9*(2), 191–226.

Lougee, B.A., & Marquardt, C.A. (2004). Earnings informativeness and strategic disclosure: An empirical examination of 'pro-forma' earnings. *The Accounting Review, 79*(3), 769–795.

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.

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research, 54*(4), 1187–1230.

Merkl-Davies, D.M., & Brennan, N.M. (2007). Discretionary disclosure strategies in corporate narratives: incremental information or impression management? *Journal of Accounting Literature, 27*(1), 116–196.

Neuendorf, K. A. (2002). The Content Analysis Guidebook. California: Sage Publications.

Rutherford, B.A. (2005). Genre analysis of corporate annual report narratives: A corpus linguistics-based approach. *The Journal of Business Communication 42*(4), 349–378.

Schleicher, T. (2010). When is good news really good news? *Accounting and Business Research, 42*(5), 547–573.

Schleicher, T., & Walker, M. (2010). Bias in the tone of forward-looking narratives. *Accounting and Business Research, 40*(4), 371–390.

Schleicher, T., & Walker, M. (2015). Are interim management statements redundant? *Accounting and Business Research, 45*(2), 229–255.

Schrand, C.M. & Walther, B.R. (2000). Strategic benchmarks in earnings announcements: The selective disclosure of prior-period earnings components. *The Accounting Review, 75*(2), 151–177.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.

Tetlock, P., Saar-Tsechansky, & M. Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437–1467.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*(2), 307–333.

Table 1
Sample Selection

Firm Sample	
Firms in FTSE All-Share Index on 30 June 2008	668
Less: Financial Firms	(305)
FTSE All-Share Index Non-Financial Firms on 30 June 2008	363
Less: Non-Financial Firms releasing Quarterly Statements in 2008	(39)
FTSE All-Share Index Non-Financial Firms disclosing IMS in 2008	**324**
Randomly Selected Non-Financial Firms from 30 June 2008	**100**
Size Composition in Selected Sample	
FTSE 100	15
FTSE 250	38
FTSE Small Cap	47
Total Firms	100
IMS Sample	
Total Number of Firms	100
Maximum Possible IMS from Sample Firms	1200
Less: Firms delisted	(69)
Less: IMS not disclosed	(109)
Final Sample of IMSs	**1022**

Notes: The table illustrates the sample selection procedure. The sampling period spans six years namely 2008–2013. 2008 is used as the year of sample determination and had 668 firms in the FTSE All-Share Index as at 30 June 2008. Eliminating all financial firms and all non-financial firms publishing full quarterly results in 2008 leaves 324 non-financial firms disclosing an IMS for 2008. We randomly select 100 firms from this list, which can yield a maximum of 1,200 IMSs during the six-year period. We subsequently lose observations due to: (i) the collapse of a firm and (ii) IMSs not being disclosed by the firm, resulting in a final sample of 1,022 IMSs during the sampling period.

Table 2
Variable Definitions

Variable	Definition	Symbol
Cumulative Abnormal Return	Return over a three-day event (days t–1, t, t+ 1) centred on the IMS release date. For abnormal returns, we calculate daily market model adjusted returns, u_{it}, as $u_{it} = R_{it} – (α_i + β_iR_{mt})$, where R_{it} is the return of firm *i* on day *t*, R_{mt} is the return of the FTSE All-Share Index on day *t* and where R_{it} and R_{mt} are calculated from DataStream Return Indices, RI. $α_i$ and $β_i$ are firm *i*'s estimated market model parameters calculated from the non-event period which runs from t–60 to t–10 and t+10 to t+60 relative to the IMS announcement day t = 0. The cumulative abnormal return is calculated as the sum of the daily market model adjusted returns, u_{it}, over the three-day event period (days t–1, t, t+1), such that $CAR_{it} = u_{it-1} + u_{it} + u_{it+1}$.	CAR
Manual Tone	The net tone score from manual scoring is computed as the difference between the number of positive and negative clauses in an IMS divided by the total number of positive and negative clauses in the IMS.	$TONE_{MANUAL}$
Automated Tone	The two automated tone scores are Henry (2008) ($TONE_{HENRY}$) and Loughran and McDonald (2011) ($TONE_{LM}$). $TONE_{HENRY}$ is the net tone score from automated scoring using the Henry (2008) wordlist and is calculated as the difference between the number of positive and negative keywords from Henry (2008) wordlist divided by the total number of positive and negative keywords. $TONE_{LM}$ is the net tone score from automated scoring using the Loughran and McDonald (2011) wordlist and is calculated as the difference between the number of positive and negative keywords from the LM wordlist divided by the total number of positive and negative keywords.	$TONE_{HENRY}$, $TONE_{LM}$
Manual Positivity and Negativity	POS_{MANUAL} is the manual positivity score and is calculated as the number of positive clauses in an IMS divided by the total number of clauses in the IMS. NEG_{MANUAL} is the manual negativity score and is calculated as the number of negative clauses in an IMS divided by the total number of clauses in the IMS.	POS_{MANUAL}, NEG_{MANUAL}
Automated Positivity and Negativity	POS_{HENRY} is the automated positivity score and is calculated as number of positive words from Henry (2008) wordlist, scaled by the total number of words in the IMS document. NEG_{HENRY} is the automated negativity score, calculated as number of negative words from Henry (2008) wordlist scaled by the total number of words in the IMS document. POS_{LM} is the automated positivity score, calculated as number of positive words from Loughran and McDonald (2011) wordlist, scaled by the total number of words in the IMS document. NEG_{LM} is the automated negativity score, calculated as number of negative words from Loughran and McDonald (2011) wordlist scaled, by the total number of words in the IMS document.	POS_{HENRY}, NEG_{HENRY}, POS_{LM}, NEG_{LM}
Firm Size	The natural logarithm of market value of the equity at the start of the financial year defined as the number of shares outstanding multiplied by the share price at the start of the financial year.	SIZE
Book-to-Market	The book value of the firm's equity at the start of the financial year divided by the market value of equity at the start of the financial year.	BTM
Profitability Status	An indicator variable taking the value of 1 if pre-exceptional operating profit is negative in the previous financial year, and zero otherwise.	LOSS
IMS Length	LENGTH[C] is the total number of clauses in the IMS document while LENGTH[W] is the total number of words in the IMS document.	LENGTH[C], LENGTH[W]

Table 3
Descriptive Statistics

Variable	OBS	Mean	Std. Dev	Median	Minimum	Maximum
CAR	1022	0.000	0.085	0.000	–0.563	1.224
TONE$_{MANUAL}$	1022	0.254	0.433	0.266	–1.000	1.000
TONE$_{HENRY}$	1022	0.588	0.290	0.636	–1.000	1.000
TONE$_{LM}$	1022	0.512	0.291	0.551	–1.000	1.000
POS$_{MANUAL}$	1022	0.245	0.143	0.229	0.000	0.824
POS$_{HENRY}$	1022	0.028	0.012	0.028	0.000	0.072
POS$_{LM}$	1022	0.044	0.018	0.043	0.000	0.112
NEG$_{MANUAL}$	1022	0.142	0.104	0.125	0.000	0.556
NEG$_{HENRY}$	1022	0.007	0.005	0.006	0.000	0.030
NEG$_{LM}$	1022	0.013	0.008	0.012	0.000	0.045
SIZE	1022	17.8	1.66	17.6	10.4	22.6
BTM	1022	0.59	1.19	0.48	–12.5	25.0
LOSS	1022	0.14	0.34	0.00	0.00	1.00
LENGTH[C]	1022	28.0	21.9	22	4	237
LENGTH[W]	1022	1010	816	777	107	9401

Notes: The table presents summary statistics of variables used in studying the full-text documents of 1,022 IMSs during the period 2008–2013. All tone measures are shown prior to standardization. Variables are as defined in Table 2.

Table 4
Distribution of Tone and CAR

Panel A: Percentiles of Tone and CAR Distributions

Percentile	TONE$_{MANUAL}$	TONE$_{HENRY}$	TONE$_{LM}$	CAR
0.01	–0.750	–0.250	–0.364	–0.244
0.05	–0.455	0.071	0.000	–0.108
0.10	–0.333	0.200	0.143	–0.070
0.25	0.000	0.455	0.333	–0.028
Median	0.273	0.636	0.551	0.000
0.75	0.556	0.793	0.724	0.034
0.90	0.867	0.900	0.850	0.068
0.95	1.000	1.000	0.923	0.095
0.99	1.000	1.000	1.000	0.220
Greater than Zero	662	965	978	511
Equal to Zero	132	7	11	17
Less than Zero	228	50	33	494

Panel B: Median Tone and CAR for subsets of the sample formed by deciles of the three tone measures

TONE$_{MANUAL}$ Deciles	TONE$_{MANUAL}$	TONE$_{HENRY}$	TONE$_{LM}$	CAR
0	–0.444	0.333	0.206	–0.009
1	–0.200	0.450	0.321	–0.014
2	0.000	0.500	0.416	–0.002
3	0.111	0.548	0.403	0.000
4	0.200	0.636	0.542	0.000
5	0.333	0.684	0.583	0.004
6	0.428	0.684	0.604	–0.001
7	0.529	0.695	0.628	0.009
8	0.750	0.798	0.738	0.011
9	1.000	0.865	0.815	0.013

TONE$_{HENRY}$ Deciles	TONE$_{MANUAL}$	TONE$_{HENRY}$	TONE$_{LM}$	CAR
0	–0.200	0.071	0.032	–0.012
1	0.000	0.333	0.250	–0.003
2	0.000	0.442	0.361	0.001
3	0.143	0.537	0.461	0.000
4	0.183	0.604	0.522	0.000
5	0.333	0.667	0.581	0.004
6	0.385	0.722	0.647	0.001
7	0.500	0.789	0.714	0.003
8	0.500	0.857	0.778	0.012
9	0.760	1.000	0.913	0.004

TONE$_{LM}$ Deciles	TONE$_{MANUAL}$	TONE$_{HENRY}$	TONE$_{LM}$	CAR
0	–0.222	0.111	0.000	–0.006
1	0.000	0.333	0.216	–0.008
2	0.016	0.471	0.333	0.001
3	0.200	0.546	0.441	–0.006
4	0.257	0.580	0.523	0.009
5	0.250	0.667	0.583	–0.007
6	0.392	0.714	0.653	0.000
7	0.500	0.778	0.724	0.014
8	0.500	0.840	0.800	0.009
9	0.800	1.000	0.923	0.007

Notes: Panel A presents the percentiles of the three tone measures distributions and the CAR distribution. Panel B reports the deciles of the three tone measures. For each decile, it takes the observations corresponding to that decile and then calculates the median value of the tone measure for which the decile is being reported, and the median values of the other tone measures and CAR for the same set of observations. All tone measures are shown prior to standardization. Variables are as defined in Table 2.

Table 5
Correlation Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13
1 = CAR	1.000												
2 = TONE$_{MANUAL}$	**0.211**	1.000											
3 = TONE$_{HENRY}$	**0.120**	**0.576**	1.000										
4 = TONE$_{LM}$	**0.149**	**0.609**	**0.889**	1.000									
5 = POS$_{MANUAL}$	**0.142**	**0.641**	**0.273**	**0.311**	1.000								
6 = POS$_{HENRY}$	**0.082**	**0.363**	**0.516**	**0.521**	**0.454**	1.000							
7 = POS$_{LM}$	**0.084**	**0.342**	**0.499**	**0.560**	**0.421**	**0.935**	1.000						
8 = NEG$_{MANUAL}$	**–0.157**	**–0.733**	**–0.526**	**–0.536**	**–0.063**	**–0.091**	**–0.092**	1.000					
9 = NEG$_{HENRY}$	**–0.124**	**–0.459**	**–0.820**	**–0.706**	–0.030	–0.025	–0.041	**0.607**	1.000				
10 = NEG$_{LM}$	**–0.143**	**–0.520**	**–0.735**	**–0.816**	**–0.087**	–0.048	**–0.063**	**0.633**	**0.874**	1.000			
11 = SIZE	0.039	**0.121**	**0.077**	**0.074**	**0.170**	**0.277**	**0.210**	–0.022	**0.090**	0.052	1.000		
12 = BTM	–0.010	**–0.076**	**–0.086**	**–0.104**	**–0.074**	**–0.062**	–0.019	0.052	**0.076**	**0.105**	**–0.236**	1.000	
13 = LOSS	0.029	**–0.081**	**–0.112**	**–0.127**	**–0.193**	**–0.158**	**–0.158**	–0.059	0.022	0.030	**–0.208**	**0.104**	1.000

Notes: The table presents Spearman's rank correlations between the discrete and continuous variables used in the study of content analysing the full-text documents of 1,022 IMSs during the period 2008–2013. The coefficients reported in bold are significant at $p < 0.05$. All tone measures are shown prior to standardization. All variables are defined as in Table 2.

Table 6
Regressions of CAR on the Manual and Automated Tone of the IMS

Panel A: CAR Regression

Variables	Net Tone Score			Positivity and Negativity		
	HENRY	LM	MANUAL	HENRY	LM	MANUAL
INTERCEPT	–0.0265	–0.0257	–0.0238	–0.0418	–0.0401	–0.0245
TONE	0.0052	0.0090**	0.0152**			
POS				0.0018	0.0027	0.0092**
NEG				–0.0123**	–0.0127**	–0.0113**
SIZE	0.0018	0.0016	0.0014	0.0025	0.0023	0.0014
BTM	0.0072	0.0072	0.0076	0.0076	0.0073	0.0080
LOSS	0.0080	0.0091	0.0100	0.0098	0.0097	0.0095
INDUSTRY FE	YES	YES	YES	YES	YES	YES
YEAR FE	YES	YES	YES	YES	YES	YES
F-VALUE	1.3	1.73**	2.98**	2.25**	2.38**	2.89**
ADJ R2	0.0049	0.0119	0.0318	0.0215	0.0237	0.0321
OBS	1022	1022	1022	1022	1022	1022

Panel B: Vuong Tests of Model Preference

	Models of Net Tone Score		
	Preferred Model	Vuong's Z-Statistic	P-value
HENRY – LM	LM	–2.96	0.003
HENRY – MANUAL	MANUAL	–2.24	0.025
LM – MANUAL	MANUAL	–2.14	0.032

	Models of Positivity and Negativity		
	Preferred Model	Vuong's Z-Statistic	P-value
HENRY – LM	NONE	–0.75	0.456
HENRY – MANUAL	MANUAL	–1.65	0.099
LM - MANUAL	NONE	–1.19	0.234

Notes: The table presents ordinary least square regressions of three-day (days t–1, t, t+1) cumulative abnormal return (CAR) on manual and automated net tone scores and positivity and negativity from content analysing the full-text documents of 1,022 IMSs from the period 2008–2013. Tone coefficients are standardized to have a mean of 0 and standard deviation of 1. Tone variables for HENRY are $\text{TONE}_{\textit{HENRY}}$, $\text{POS}_{\textit{HENRY}}$ and $\text{NEG}_{\textit{HENRY}}$ respectively. Tone variables for LM are $\text{TONE}_{\textit{LM}}$, $\text{POS}_{\textit{LM}}$ and $\text{NEG}_{\textit{LM}}$ respectively. Tone variables for MANUAL are $\text{TONE}_{\textit{MANUAL}}$, $\text{POS}_{\textit{MANUAL}}$ and $\text{NEG}_{\textit{MANUAL}}$ respectively. Coefficients marked with (***) are significant at p < 0.01. Coefficients marked with (**) are significant at p < 0.05. Coefficients marked with (*) are significant at p <0.1. F-VALUE: model F-statistic. OBS: number of observations. Coefficient p-values are based on two-way clustered standard errors. Clustering is performed by firm and year. All other variables are defined in Table 2.

Table 7
Analysis of the Differences between Manual and Automated Tone for Positive and Negative CAR

		NUMBER OF CLAUSES						
			Incidence of Keywords (KW)		Clauses with Multiple Keywords		Clauses with Non-Performance Context	
		Clauses	Clauses With 0 KW	Clauses With >0 KW	Multiple Positive KW	Multiple Negative KW	With Positive KW	With Negative KW
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
CARNEG –	HENRY	210	123	87	26	7	21	20
HIGHMAN					29.9%	8.0%	24.1%	23.0%
	LM	200	62	138	33	16	28	24
					23.9%	11.6%	20.3%	17.4%
CARPOS –	HENRY	159	90	69	24	2	8	16
HIGHMAN					34.8%	2.9%	11.6%	23.2%
	LM	296	98	198	36	25	32	49
					18.2%	12.6%	16.2%	24.7%
TOTAL –	HENRY	369	213	156	50	9	29	36
HIGHMAN					32.1%	5.8%	18.6%	23.1%
	LM	496	160	336	69	41	60	73
					20.5%	12.2%	17.9%	21.7%
CARNEG –	HENRY	334	176	158	38	7	24	13
LOWMAN					24.1%	4.4%	15.2%	8.2%
	LM	340	128	212	53	36	65	24
					25.0%	17.0%	30.7%	11.3%
CARPOS –	HENRY	353	175	178	44	19	33	11
LOWMAN					24.7%	10.7%	18.5%	6.2%
	LM	279	94	185	46	20	57	33
					24.9%	10.8%	30.8%	17.8%
TOTAL -	HENRY	687	351	336	82	26	57	24
LOWMAN					24.4%	7.7%	17.0%	7.1%
	LM	619	222	397	99	56	122	57
					24.9%	14.1%	30.7%	14.4%
TOTAL	HENRY	1056	564	492	132	35	86	60
					26.8%	7.1%	17.5%	12.2%
	LM	1115	382	733	168	97	182	130
					22.9%	13.2%	24.8%	17.7%

Notes: The table presents the differences between manual and automated tone in the clauses of IMSs associated with positive and negative CAR values. TONE_{MANUAL} is regressed in separate models on TONE_{HENRY} and TONE_{LM}. In each case, 10 IMSs with the largest positive model residuals and 10 IMSs with the largest negative model residuals are identified when CAR > 0 and CAR <=0. This gives, for each of HENRY and LM, in total 40 IMSs with the largest positive and negative differences with manual scoring when CAR > 0 and CAR <=0. Then for these IMSs, we present the total number of clauses, the number of clauses containing no keyword, the number of clauses containing at least one key word, the number of clauses containing multiple positive and multiple negative keywords, and the number of clauses in a non-performance context containing positive or negative keywords. The percentages in columns (4) to (7) represent the number of clauses reported in that column divided by the number of clauses reported in column (3). CARPOS means CAR > 0. CARNEG means CAR <=0. HIGHMAN = ten largest positive regression residuals between manual and automated tone. LOWMAN = ten largest negative regression residuals between manual and automated tone.

Table 8
Regressions of CAR on the Manual and Automated Tone of the IMS Partitioned by Firm Size

Panel A: CAR Regression

Variables	Net Tone Score (Large Firms)			Net Tone Score (Small Firms)		
	HENRY	LM	MANUAL	HENRY	LM	MANUAL
INTERCEPT	0.0517	0.0460	0.0450	–0.0064	0.0174	0.0439
TONE	0.0035	0.0068**	0.0051	0.0069	0.0110*	0.0237***
SIZE	–0.0021	–0.0018	–0.0018	–0.0004	–0.0019	–0.0032
BTM	–0.0031	–0.0029	–0.0031	0.0129	0.0126	0.0139
LOSS	–0.0004	0.0009	–0.0002	0.0143	0.0145	0.0173
INDUSTRY FE	YES	YES	YES	YES	YES	YES
YEAR FE	YES	YES	YES	YES	YES	YES
F-VALUE	1.76**	2.04***	1.93**	1.18	1.38	2.51***
ADJ R2	0.0228	0.0308	0.0278	0.0062	0.0130	0.0495
OBS	556	556	556	466	466	466

Panel B: Vuong Tests of Model Preference

	Large Firms		
	Preferred Model	Vuong's Z-Statistic	P-value
HENRY – LM	LM	–1.85	0.065
HENRY – MANUAL	MANUAL	–1.74	0.081
LM – MANUAL	NONE	0.35	0.724

	Small Firms		
	Preferred Model	Vuong's Z-Statistic	P-value
HENRY – LM	NONE	0.48	0.635
HENRY – MANUAL	MANUAL	–2.65	0.008
LM – MANUAL	MANUAL	–2.31	0.021

Notes: The table presents ordinary least square regressions of three-day (days t–1, t, t+1) cumulative abnormal return (CAR) on manual and automated net tone scores of full-document IMSs from the period 2008–2013, partitioned by firm size. The large firm sample consists of 556 IMSs of FTSE100 and FTSE250 constituents and the small firm sample consists of 466 IMSs of FTSE SmallCap constituents. Tone coefficients are standardized to have a mean of 0 and standard deviation of 1. Tone variable for HENRY is $TONE_{HENRY}$, for LM is $TONE_{LM}$ and for MANUAL is $TONE_{MANUAL}$. Coefficients marked with (***) are significant at p < 0.01. Coefficients marked with (**) are significant at p < 0.05. Coefficients marked with (*) are significant at p <0.1. F-VALUE: model F-statistic. OBS: number of observations. Coefficient p-values are based on two-way clustered standard errors. Clustering is performed by firm and year. All other variables are defined in Table 2.

Appendix A

Manual and Automated Tone Scoring: Examples

Example 1. *Group earnings have been in line with the Board's expectations of flat trading in the first half.* (Cobham plc, IMS published on 10.7.2013)
MANUAL Neutral, 0. HENRY 0, 0, 0. LM 0, 0, 0.

Example 2. *We anticipate full year commodity costs to increase by approximately 20 million,* [together with a 12 million increase in utilities]. (Northern Foods plc, IMS published on 28.7.2008)
MANUAL Negative, -1. HENRY 1 (increase), 0, 1. LM 0, 0, 0.

Example 3. *Upstream profits are adversely impacted by lower commodity prices,* [reducing total Group operating profit]. (Centrica plc, IMS published on 11.5.2009)
MANUAL Negative, -1. HENRY 0, 1 (lower), -1. LM 0, 1 (adversely), -1.

Example 4. *The Group's business depends on good relations with its employees and with the communities surrounding its operations.* (Hochschild Mining, IMS published on 7.10.2009)
MANUAL Neutral, 0. HENRY 1 (good), 0, 1. LM 1 (good), 0, 1.

Example 5. *We now estimate that the total unit shipment for this financial year will exceed 75 million units, which compares to the 47 million units shipped for last financial year.* (Imagination Technologies, IMS published on 19.3.2009)
MANUAL Positive, 1. HENRY 1 (exceed), 0, 1. LM 0, 0, 0.

Example 6. [Costs continue to be tightly managed], *like for like costs in the period were 5.2% below last year (against a 7.8% decrease in the six month period to 30 September 2009).* (BSS Group plc, IMS published on 11.2.2010)
MANUAL Positive, 1. HENRY 0, 2 (below, decrease), -1. LM 0, 0, 0.

Example 7 *For the financial year 2010/11 we expect underlying profit before tax to be around market consensus.* (Shanks Group, IMS published on 4.2.2011)
MANUAL Neutral, 0. HENRY 0, 0, 0. LM 0, 0, 0.

Example 8. *Non key account revenues grew 6%.* (Brammer plc, IMS published on 17.5.2012)
MANUAL Positive, 1. HENRY 1 (grew), 0, 1. LM 0, 0, 0.

Example 9. *We continue to forecast an underlying Group effective tax rate of around 55% due to the high proportion of upstream profits.* (Centrica plc, IMS published on 31.10.2008)
MANUAL Neutral, 0. HENRY 1 (high), 0, 1. LM 1 (effective), 0, 1.

Example 10. *The outlook for 2011* [indicates that conditions are no longer worsening and], [combined with the substantial cost cuts successfully implemented during the summer as announced at the half year], *leads us to a more optimistic view for next year.* (Intec Telecom Systems, IMS published on 19.8.2010)
MANUAL Positive, 1. HENRY 0, 0, 0. LM 1 (optimistic), 0, 1.

Example 11. [Despite the increase in active customers], *trading in this online division has been softer than anticipated.* (Sportech plc, IMS published on 19.11.2009)
MANUAL Negative, -1. HENRY 0, 0, 0. LM 0, 0, 0.

Example 12. *Including disposed businesses, group revenue was up 14% year to date.* (Associated British Foods plc, 9.7.2009)
MANUAL Positive, 1. HENRY 1 (up), 0, 1. LM 0, 0, 0.

Example 13. *Consumer demand remained muted during November and early December.* (Hornby plc, IMS published on 27.1.2009)
MANUAL Negative, -1. HENRY 0, 0, 0. LM 0, 0, 0.

Notes: This appendix presents the difference between manual and automated tone scores of some selected clauses. Company names and IMS publication dates are given in parenthesis ( ) after the clause. The clauses scored are presented in italics. Separate clauses within a textual sentence, the scores of which are not shown, are separated with brackets [ ]. Guide for reading scores – MANUAL: Tone, Clause Tone Score. HENRY (Automated): Positive words, Negative words, Clause Tone Score. LM (Automated): Positive words, Negative words, Clause Tone Score.

Appendix B
Comparison of Manual and Automated Tone Assignment: Examples

Negative Words	List(s)	Count	Clauses	POS_{MANUAL}	NEU_{MANUAL}	NEG_{MANUAL}	Agreement
Below	Henry	432	389	62	173	154	39.59%
Weak	Henry	497	496	121	182	193	38.91%
Decreased	Henry	165	153	22	21	110	71.90%
Decrease	Henry	61	61	9	22	30	49.18%
Uncertainty	Henry	222	222	26	55	141	63.51%
Uncertain	Henry	92	92	11	48	33	35.87%
Difficult	Henry	381	381	76	87	218	57.22%
Lower	Henry	910	851	203	182	466	54.76%
Declined	Both	263	262	31	28	203	77.48%
Decline	Both	376	358	71	109	178	49.72%
Challenging	Both	539	531	97	217	217	40.87%
Challenges	Both	132	132	14	85	33	25.00%
Losses	LM	50	50	10	25	15	30.00%
Loss	LM	121	80	9	39	32	40.00%
Against	LM	407	385	119	185	81	21.04%
Claims	LM	14	14	1	9	4	28.57%
Disclosed	LM	41	31	5	17	9	29.03%
Adversely	LM	48	48	2	20	26	54.17%
Impairment	LM	27	27	0	23	4	14.81%
Adverse	LM	178	178	30	84	64	35.96%
Termination	LM	104	102	16	47	39	38.24%
Litigation	LM	28	28	3	12	13	46.43%

Positive Words	List(s)	Count	Clauses	POS_{MANUAL}	NEU_{MANUAL}	NEG_{MANUAL}	Agreement
Increased	Henry	1314	1245	748	350	147	60.08%
Increase	Henry	998	954	442	368	144	46.33%
Grew	Henry	435	424	350	26	48	82.55%
Growth	Henry	3424	3394	1247	1735	412	36.74%
Exceed	Henry	104	102	58	36	8	56.86%
Rise	Henry	191	191	49	110	32	25.65%
Deliver	Henry	531	520	120	360	40	23.08%
Up	Henry	1431	1218	729	351	138	59.85%
Strong	Both	2678	2552	1077	1227	248	42.20%
Good	Both	1279	1239	580	550	109	46.81%
Greater	Both	208	179	29	128	22	16.20%
Best	Both	124	124	28	91	5	22.58%
Improvements	Both	181	181	67	91	23	37.02%
Opportunities	Both	648	629	82	504	43	13.04%
Effective	LM	124	123	40	58	25	32.52%
Benefit	LM	535	526	171	302	53	32.51%
Gain	LM	272	271	90	148	33	33.21%
Gains	LM	534	516	170	248	98	32.95%
Able	LM	103	103	10	86	7	9.71%
Advances	LM	15	15	10	5	0	66.67%
Successful	LM	452	447	84	351	12	18.79%
Beneficial	LM	23	23	9	12	2	39.13%
Favorable	LM	125	125	54	43	28	43.20%

Notes: This appendix presents some high frequency negative and (non-negated) positive words from the automated wordlists and the manual tone assigned to the clauses in which these words appear. 'Count' is the total incidence of these words in 1,022 IMSs, 'Clauses' is the total number of clauses they appear in, and 'Agreement' is the percentage of agreement / similarity between the manual and automated methods with regards to the tone assignment. NEU_{MANUAL} indicates clauses that were neither positive nor negative in the manual scoring.