Textual sentiment in finance: A survey of methods and models

Colm Kearney and Sha Liu

Research Paper
April 2013

Abstract

We survey the textual sentiment literature, comparing and contrasting the various information sources, content analysis methods, and empirical models that have been used to date. We summarize the important and influential findings about how textual sentiment impacts on individual, firm-level and market-level behavior and performance, and *vice versa*. We point to what is agreed and what remains controversial. Promising directions for future research are emerging from the availability of more accurate and efficient sentiment measures resulting from increasingly sophisticated textual content analysis coupled with more extensive field-specific dictionaries. This is enabling more wide-ranging studies that use increasingly sophisticated models to help us better understand behavioral finance patterns across individuals, institutions and markets.

Keywords:
Behavioral finance, textual sentiment, internet messages, news, market efficiency

JEL Classification:
D80; D82; G02; G10; G12; G14; G30; G34;G38;M41

Contact details
Colm Kearney, Faculty of Business and Economics, Monash University, Sir John Monash Drive, Melbourne, Australia. Tel: +61 3 99051021, Email: colmpkearney@gmail.com
Sha Liu, Business School, Trinity College Dublin, College green, Dublin 2, Ireland. Email: lius2@tcd.ie.

1 Introduction

Almost half a century ago, Stone *et al*. (1966) described how words and sentences are quintessential human artifacts, the products of social constructs and experiences that provide essential evidence about individual and collective processes such as economic and financial activities and behaviors. They defined content analysis as any technique that enables inference by objectively and systematically identifying specified characteristics within text. By analyzing the raw data of words and sentences, behavioral scientists have become increasingly skilled at uncovering the evidence of sentiment or affect within text. Sentiment is now understood to be articulated in many forms of human discourse, public speeches, news reports, blogs and other forms of written, spoken and visual communication.

In behavioral finance, researchers have in the past decade intensified their efforts to understand how sentiment impacts on individual decision-makers, institutions and markets. Broadly speaking, two types of sentiment have been studied. The first is investor sentiment – beliefs about future cash flows and investment risks that are not justified by the facts at hand (Baker and Wurgler (2007)). A substantial body of this literature focuses on finding and quantifying the effects of investor sentiment on individual stocks and the overall market using various ways to measure investor sentiment. The second type of sentiment is text-based or textual sentiment – the degree of positivity or negativity in texts. In some studies, particularly those using corporate disclosures as the information source, the term 'tone' (positive or negative) is used to refer to sentiment. In broad terms, however, textual sentiment may also include affects other than positivity-negativity, such as strong-weak, and active-passive.

The fundamental difference between investor sentiment and textual sentiment is that the former captures the subjective judgments and behavioral characteristics of investors, while the latter can include the former, but also includes the more objective reflection of conditions within firms, institutions and markets. The connection between textual sentiment and investor sentiment is complex, and the extent to which they are causally related has not yet been

thoroughly examined or understood. It is also unclear how investors interpret textual sentiment. The existing studies tend not to make assumptions about investor rationality, or about the relationship between textual sentiment and investor behavior. In this sense, they transcend the boundaries between behavioral and traditional finance. The inclusion of qualitative information from textual sentiment into equity asset pricing models, for example, provides another perspective and potentially complementary information to quantitative informational measures in the price formation process. Qualitative information in publicly available documents or media articles may contain additional hard-to-quantify information. Li (2006) suggests that text-based information can potentially provide a more independent test of market efficiency than the number-based measures, because many of the latter are highly correlated so different anomalies may reflect the same empirical regularity. More generally, however, textual sentiment analysis provides an increasingly important approach to address many pivotal questions in behavioral finance.

In this paper, we review the burgeoning literature that uses textual analysis to extract sentiment from sources such as corporate disclosures, media articles, and internet postings[1]. We describe the alternative content analysis methods including the dictionary-based approaches and the machine learning techniques that are commonly used to generate the sentiment series. Corporate disclosure studies usually aim to discover the fundamental relation between sentiment and future firm performance or other quantitative variables. Media articles and internet posting studies focus on the short-term effects of sentiment on market variables such as stock prices, returns, trading volumes and volatility. Each information source and linguistic method has its unique advantages and disadvantages which influence the research focus and limitations of each study. Because of the use of different types of sentiment and varied research focuses, a wide variety of models and methods has been used to test alternative hypotheses and to derive new findings.

Our paper is structured as follows. Section 2 describes and discusses the qualitative information sources used in the literature. Each of the most popular information sources –

public corporate disclosures, news articles and internet messages – are described along with their unique features that are advantageous to others. Section 3 introduces and compares the most frequently used textual analysis methods: the dictionary-based approach and the machine learning approach. Sentiment measures derived from the different linguistic analysis methods are introduced in Section 4. Section 5 presents and reviews the empirical models that have been used in the literature to date, including contemporaneous linear regressions, univariate and multivariate time-series models, logistic regressions and volatility models. The general forms of models are summarized. This section also describes the ex-post sentiment-based trading strategies that are discussed in the literature. Section 6 summarizes the main findings of the literature to date. Section 7 summarizes the paper, draws together the most important conclusions and suggests future research directions.

2 Information sources

The qualitative information that has been analyzed by textual sentiment researchers in finance comes predominantly from three sources: public corporate disclosures/filings, media articles and internet messages. The sentiment expressed in these texts conveys market participants' and commentators' information and opinions about many aspects of developments in firms, institutions and markets. It also reflects how sentiment responds to these events. The most important work on sentiment analysis within corporate disclosures/filings includes Li (2006), Feldman et al (2008), Henry(2008), Henry and Leone (2009), Li (2010), Davis et al. (2011), Davis and Tama-Sweet (2011), Demers and Vega (2011), Doran et al. (2010), Huang et al. (2011), Loughran and McDonald (2011a, 2011b), Rogers et al. (2011), Davis et al. (2012), Ferris et al. (2012), Jegadeesh and Wu (2012), Price et al. (2012), and Loughran and McDonald (2013a). The literature that extracts sentiment from news articles or analyst's reports includes Tetlock (2007), Engelberg(2008), Tetlock et al. (2008), Sinha (2010), Carretta et al. (2011), Engelberg et al. (2012), Ferguson et al. (2012), Garcia (2012), Rees and Twedt (2012), and Huang et al. (2013). A small group of studies examines internet-posting sentiment, foremost amongst them being Antweiler and Frank (2004), Das and Chen (2007)

and Chen et al. (2013). A summary of the information sources used in each of these papers (and papers that with a minor focus on sentiment analysis) is presented in Table 1. We now discuss each information source in turn.

Corporation-expressed sentiment

Public corporate disclosures are a natural source of textual sentiment for researchers insofar as they are official releases that come from insiders who have better knowledge of the firm than outsiders. The linguistic style and tone in these texts might convey useful information about expected future firm performance over and above the numbers in financial statements. This corporation-expressed sentiment is particularly useful in examining the role of qualitative information about individual firm performance and stock pricing. Offsetting this, however, is the low frequency of the data, because firms usually make these disclosures on a quarterly or annual basis.

The corporation-expressed sentiment literature mainly studies corporate annual or interim reports, or earnings press releases and earnings conference calls. The former category includes Li (2006), Feldman et al (2008), Li (2010), Loughran and McDonald (2011a, 2011b), and Jegadeesh and Wu (2012). In particular, Feldman et al. (2008) and Li (2010) focus only on the MD&A sections of 10-Ks and 10-Qs. Loughran and McDonald (2011a) also use MD&A texts in one part of their analysis. The earnings disclosures category includes Henry (2006), Henry(2008), Henry and Leone (2009), Doran et al. (2010), Davis et al. (2011), Davis and Tama-Sweet (2011), Demers and Vega (2011), Huang et al. (2011), Davis et al. (2012), Larcker and Zakolyukina (2012), and Price et al. (2012). In these studies, textual sentiment is considered as a new factor in addition to the usual firm-level fundamentals that are analyzed in typical event studies. Feldman et al. (2008) show how MD&As might be a superior information source relative to earnings announcements because the preliminary earnings announcements were typically not filed with the SEC prior to 2003, hence not routinely scrutinized as periodic reports with the SEC. Moreover, MD&As contain more qualitative information from management than earnings announcements, because the purpose of MD&A

is to provide a management perspective within the narrative on their firms' past performance, their current financial positions, and their future prospects.

Media-expressed sentiment

Media-expressed sentiment is the positivity or negativity contained in news stories, in-depth commentaries or analyst reports. These texts are relevant to general economic conditions, to overall financial market conditions and prospects, and to individual industries and firms. Obviously, sentiment in media articles about general economic conditions or financial markets is an appropriate choice for studying market-wide price patterns and activities, while industry-specific or firm-specific news stories and analyst reports are more appropriate for analyzing firm-level stock prices, returns, trading volumes and other performance indicators and attributes.

In the media-expressed sentiment literature, Tetlock (2007) and Garcia (2012) study general economics and finance news in two major U.S. newspapers, the *Wall Street Journal* and the *New York Times*, respectively. Engelberg (2008), Tetlock et al. (2008), Sinha (2010), and Engelberg et al. (2012) employ much wider news sources, and they all research firm-specific news. Engelberg (2008) collects news articles addressing earnings announcements of 4,700 unique firms; Tetlock et al. (2008) and Engelberg et al. (2012) use the *Dow Jones News Service* in addition to the *Wall Street Journal*; and Sinha (2010) analyzes firm-level sentiment in 587,719 news articles provided by the *Thomson Reuters NewsScope* service. Using wider news sources rather than a small number of particular newspapers is generally more appropriate and preferred, because it decreases the probability of getting biased sentiment from some sources that do not represent the overall market. Ferguson et al. (2012) is the first study that extracts firm-specific sentiment for non-US stocks. Their sample consists of 264,647 firm-specific UK news media articles from *The Financial Times, The Times, The Guardian and Mirror*, covering FTSE 100 firms.

Rees and Twedt (2012) and Huang et al. (2013) are the only researchers to date that have examined sentiment in analyst reports relating to specific firms. Rees and Twedt (2012) argue that financial analysts perform a pivotal role in evaluating accounting data and disseminating their analysis to the public, and that this ensures their continued interest from investors and academic researchers.

Internet-expressed sentiment

Internet postings are a potentially useful source of textual sentiment because many people spend a considerable amount of time every day reading and writing internet postings about stocks. The message flows comprise potentially valuable insights, market sentiment, manipulative behavior, and reactions to other sources of news (Das and Chen (2007)), possibly causing a significant impact on financial markets. Internet-expressed sentiment is potentially 'noisier' than corporation-expressed or media-expressed sentiment, because it contains more views from individual traders. This can make it a potentially powerful source from which to extract small investor sentiment. Even if a high proportion of these messages contain noise or irrational sentiment, Black (1986) argues that the narratives around a large number of small events could generate more potent causal sentiment factors than the narratives around a smaller number of large events if the former have greater proportions of noise within them.

Antweiler and Frank (2004), Das and Chen (2007) and Chen et al. (2013) have conducted computational textual sentiment analysis on internet messages. Antweiler and Frank (2004) analyzes 1.5 million messages posted on *Yahoo!Finance*[2] and *Raging Bull*[3] about the 45 companies in the *Dow Jones Industrial Average* index and the *Dow Jones Internet Index*. Das and Chen (2007)'s sample consists of all messages posted on the *Yahoo!* message boards from July to August 2001 that relate to the 24 technology-sector stocks presented in the *Morgan Stanley High-Tech Index* (MSH), resulting in a total of 145,110 messages. Chen et al. (2013) analyze 79,142 single-ticker articles published between 2005 and 2011 and their commentaries written on the *Seeking Alpha*[4] website.

Comparing and contrasting the three main sentiment sources

The advantages and disadvantages of each information source are sixfold. *First*, to the extent that corporate disclosures convey sentiment from management – the insiders who know most about their firms – they are a potentially valuable textual source. But it is unlikely that management reveals 'the truth, the whole truth, and nothing but the truth', and it may be tempted to seek to manipulate investors' judgments. Because these narratives are firm-specific, they are particularly relevant to studying the role of qualitative information in individual firm performance and stock pricing. In contrast, textual sentiment from news stories can be used in the context of either market-level or firm-level analysis depending on the subject matter of the news stories. So although news stories are a more flexible information source than corporate disclosures, sentiment from outsiders is less likely to capture insiders' views and perspectives. The qualitative information in analyst reports are likely to sit somewhere between corporate disclosures and news stories. Analysts may have some insider information that is communicable to investors and other market participants and commentators through sentiment.

Second, news stories cover many different kinds of events. From the event-study perspective, researchers can potentially examine the impact of a whole range of events on firm earnings and stock returns. In this case news, stories are superior to corporate disclosures insofar as the latter contain limited event types such as earnings conference calls and annual reports. Tetlock et al. (2008) argue that 'examining all newsworthy events simultaneously limits the scope for dredging for anomalies' (Fama (1998): 1438).

Third, news stories predominantly reflect hindsight rather than foresight. They are usually written about what has happened or what is scheduled to happen rather than what might happen in the future. In contrast, corporate disclosures such as the MD&A sections of 10-Ks, the transcripts of conference earnings calls, and analyst reports tend to contain more

forward-looking statements. Sentiment in these narratives has more potential predictability on future outcomes such as firm performance and stock prices or returns.

Fourth, because corporate releases are usually on a quarterly or annual basis, the sentiment series extracted from them is not ideal for time-series modeling because of its low frequency. It is, however, appropriate for cross-sectional analysis and event studies that examine the effect of sentiment on stock returns around corporate releases. Once again, media articles are a more flexible information source for extracting textual sentiment that can be used in modeling monthly, weekly, daily and even intra-day levels, as long as the relevant texts are available at the required frequencies.

Fifth, because the online media tends to be open and unregulated, internet-expressed sentiment is likely to be noisier than corporation-expressed or media-expressed sentiment. It is also likely to contain little new information that is incremental to published public news. A large proportion of internet messages are written and posted by noise traders or uninformed investors who might be susceptible to particular opinions and sentiments, and their 'information' is likely to be less accurate or reliable. In other words, their comments and advice may have a low ratio of information to noise. It follows that internet postings are not an ideal source of information for testing market efficiency. Insofar as they better represent small investor sentiment, they might provide more evidence for the behavioral finance framework.

Sixth, the pre-processing of internet postings prior to content analysis is more costly than corporate disclosures and media articles, because people tend to write less accurately, clearly and formally on internet boards, and the meaning of the texts can be ambiguous. Corporate documents and media articles are written more professionally and require less pre-processing time. This is probably the reason why internet postings have been the least frequently studied source of textual sentiment analysis. Since each of the information sources has its advantages and disadvantages, the best practice may be to employ as many information sources as

possible. Kothari et al. (2009) investigate an extensive set of sources from the print media, including a combination of corporate disclosures, analyst reports and news stories in financial press. Their sample comprises all available disclosure texts for 887 firms, amounting to 326,357 texts in total.

3 The methods of content analysis

The most common content analysis methods in the textual sentiment literature are the *dictionary-based approach* and *machine learning*. Table 2 summarizes the methodology employed in each study in the textual sentiment literature in finance.

The dictionary-based approach

The dictionary-based approach uses a mapping algorithm in which a computer program reads text and classifies the words, phrases or sentences into groups based on pre-defined dictionary categories (Li (2010)). This is often referred to as the 'bag-of-words' model in natural language processing. Documents are considered to be the bag of words, and the structure along with any linear ordering of words within the context is ignored (Manning and Schutze 1999). There are two important issues in the dictionary-based approach. The first is the word lists, i.e. the words contained in each sentiment category; and the second is how each word in the word lists should be weighted. We discuss each of these in turn.

Dictionaries and word lists

The most popular word lists in the earlier studies are the General Inquirer (GI) built-in dictionary developed and used by Philip Stone, a specialist in social psychology (Stone et al. (1966)). Most of its word lists come from the *Harvard IV-4* dictionaries.[5] and examples of GI word lists. A number of financial researchers have used the GI program and word lists (or the GI/Harvard word lists only) to derive sentiment in texts, including Tetlock (2007), Engelberg (2008), Feldman et al. (2008), Tetlock et al. (2008), Henry and Leone (2009), Kothari et al. (2009), Doran et al. (2010), Carretta et al. (2011), Demers and Vega (2011), Loughran and

McDonald (2011a), Engelberg et al. (2012), Ferris et al. (2012), Ozik and Sadka (2012), Price et al. (2012), and Rees and Twedt (2012).

Another frequently used dictionary is built from the popular textual analysis program, *DICTION* (Hart (2000)), which was developed by Roderick Hart, a specialist in politics and mass media. *DICTION 5.0* counts words based on 33 separate dictionaries and two sets of variables, and produces outputs of raw frequencies, percentages, and standardized scores for these lists of words. Included with the dictionary scores are scores for five master variables - *activity*, *certainty*, *commonality*, *optimism* and *realism* - and other calculated variables. A description of the *DICTION 5.0* software from its official website[6] is attached in Appendix 3. Henry and Leone (2009), Davis et al. (2011), Davis and Tama-Sweet (2011), Demers and Vega (2011), Durnev and Mangen (2011), Rogers et al. (2011), Davis et al. (2012), and Ferris et al. (2012) all employ *DICTION* in their research.

It is important to note that both the *GI* and *DICTION* are general English language linguistic dictionaries rather than dictionaries that are specific to the domain of financial disclosure. Li (2010) finds that the classification of tone based on the *GI/Harvard* does not provide sufficient accuracy. This is not surprising, because general word lists omit words that are considered positive or negative in the context of financial disclosure, and include other words that would not (Henry and Leone (2009)). For example, words such as 'tax' and 'liability' are on the negative wordlist of the *GI/Harvard*, but are not negative in the financial context. Loughran and McDonald (2011a) find that almost three-fourths (73.8 percent) of the negative word counts in the *GI/Harvard* list are attributable to words that are typically not negative in a financial context. To overcome this problem, dictionaries / word lists specific to the finance domain have been built by researchers so that more accurate and efficient sentiment scores can be generated. Studies that focus on comparing various dictionaries include Henry and Leone (2009) and Loughran and McDonald (2011a). The former uses three word lists – *DICTION*, *GI/Harvard* and Henry (2006, 2008) – to gauge the tone of earnings press releases. The authors find that the context-specific word list developed by Henry (2006, 2008) is more

powerful than the general word lists used in previous research. Later research by Price et al. (2012) also concludes that the Henry (2008) dictionary is more powerful in detecting cumulative abnormal returns beyond the initial reaction window. Loughran and McDonald (2011a) use *GI/Harvard* negative words which have been expanded by inflecting each word to forms that retain the original meaning of the root word, and the finance-specific word lists developed by themselves (L&M lists) to assess sentiment in 10-Ks[7]. They argue that although the apparent power of the two negative word lists is similar, the use of the finance-specific list to avoid those words in the *GI/Harvard* list that might proxy for industry or other unintended effects. The L&M lists have become predominant in more recent studies by Doran et al. (2010), Huang et al. (2011), Ferguson et al. (2012), Garcia (2012), Jegadeesh and Wu (2012), Chen et al. (2013), Liu and McConnell (2013), and Loughran and McDonald (2013a).

Term weighting

Most studies employ proportional weighting, which treats every word in the list to be equally important. Typical examples are the studies that use the *GI/Harvard*, which calculates simple frequencies for words appearing in the text that fall within each category. Term weighting, however, allows for the possibility that raw word counts are not the best measure of a word's information content (Loughran and McDonald (2011a)). Loughran and McDonald (2011a) use two weighting schemes, a simple proportional weighting and one that weights each word inversely proportional to its document frequency, or the frequency with which each it appears in the sample of documents. Jegadeesh and Wu (2012) argue that in the context of inferring tone, there is no reason to expect why words that are found in fewer documents should be more powerful than those found in more documents. These researchers use a weighting scheme that is particularly suitable for finance applications. They assign weights for each word based on how the market has reacted to them in the past. Their results indicate that the appropriate choice of term weighting is more important than a complete and accurate dictionary to which the weighting scheme is applied.

Figure 1 shows a generic process of using the dictionary-based approach to extract sentiment from texts. The first step is to collect qualified texts to form the corpus. Pre-processing of the collected files is usually necessary in order to split the articles and group them by date. Subsequently, the sentiment categories (i.e. the word lists) in the content analysis program should be selected. Researchers can also write their own programs which allow the addition of customized word lists. Sentiment scores are obtained by selecting the text files and running the program. It is also possible to construct other sentiment measures based on the original sentiment scores. Finally, the sentiment measures and other variables can be used for financial modeling and hypothesis testing.

Machine learning

Pioneered by mathematicians and computer scientists, machine learning relies on statistical techniques to infer the content of documents and to classify them based on statistical inference (Li (2010)). The steps involved in this method are as follows. A proportion of the complete corpus of text to be analyzed is designated as the 'training set'. Each word in the training set is manually classified as 'positive', 'negative', or some other dimension of sentiment (such as 'strong', 'weak', 'active', 'passive'). A selection of sentiment analysis algorithms (e.g. the Naïve Bayesian algorithm) is then trained on the training corpus. The algorithms 'learn' the sentiment classification rules (or 'grammar') from the pre-classified data set, and apply these rules out-of-sample to the whole corpus. When all words in the complete corpus are classified, sentiment measures can be derived using various combinations of the classifications used in the training corpus. In essence, machine learning involves one or more algorithms reading a training set and writing a 'model' containing its statistics, which is then applied to the whole corpus to derive textual sentiment scores.

Although machine learning can be carried out by customized programs, two established programs are worth mentioning. The first is the *Rainbow* package developed by McCallum (1996). *Rainbow* supports several alternative classification methods. The default is *Naive Bayes*, but k-nearest neighbor, *TFIDF*, and *probabilistic indexing* are all available. Another

text-processing tool is the *Reuters NewsScope Sentiment Engine*[8]. This engine has three major processes: pre-processing; the lexical and sentiment pattern identifier; and the sentiment classifier. In the first two processes, documents are split into sentences which are further split into words; the subject of the sentence is identified; and each word is identified as a noun, verb, adjective, adverb, and intensifier. The last process, sentiment classification, is done by a three layer back-propagation neural network. It follows the general steps of machine learning, and produces as output the probability of the text being positive, negative, or neutral.

Antweiler and Frank (2004) use the default method *Naïve Bayesian* algorithm within the *Rainbow* package to classify messages into one of three categories: *buy, sell* and *hold*. The *Naïve Bayesian* algorithm is also employed by Li (2010) to classify sentences into one of four tones: *positive, neutral, negative,* and *uncertain*. Das and Chen (2007) use five algorithms to classify internet messages into *bullish, bearish* or *neutral*. Sinha (2010) uses the *Reuters NewsScope Sentment Engine* to calculate the probabilities of news articles being positive, negative and neutral, respectively. Huang et al. (2013) use the Naïve Bayes machine learning approach to extract textual opinions from analyst reports about *S&P 500* firms.

Comparing and contrasting the textual analysis methods

The dictionary-based and machine learning approaches have their advantages and disadvantages. *First*, the dictionary-based approach is probably the easiest for business, economic and financial analysts to handle because the well-established programs like *GI* and *DICTION* are readily available and have been most frequently used in the literature. As discussed previously, however, the general dictionaries are not appropriate for textual analysis in financial contexts. This problem can be largely mitigated by applying finance-specific word lists such as the *L&M* lists, and the key issue with this methodology lies in the choice of the most appropriate weighting scheme. The latter most likely depends on the nature of the corpus being analyzed and the hypotheses being tested.

Second, implementing machine learning is more time-consuming and costly than the dictionary-based approach, because the text in the 'training set' must be manually classified. In addition, to ensure the highest quality of classification, the criteria for selecting appropriate people to read the texts are strict (e.g. native speakers with economic and finance backgrounds). The accuracy rate of machine learning is usually higher than the dictionary-based approach (Li (2010)). For example, the *Reuters NewsScope Sentiment Engine* boasts a 75 percent accuracy rate compared to the average assessment of human analysts (Sinha (2010)). Huang et al. also document that their classification accuracy achieved using the Naïve Bayes machine learning approach is 80.9% in the in-sample validation and 76.9% in the out-of-sample validation, which is substantially higher than that achieved using the dictionary-based approaches based on the general dictionaries (e.g. 48.4% for *GI* and 54.9% for *DICTION*).

4 Measuring textual sentiment

Having extracted sentiment from a corpus of text, the construction of sentiment measures is relatively straightforward. Due to differences in the methods of content analysis, however, the characteristics of the resulting sentiment measure are correspondingly diverse. We now summarily describe and compare the most common measures of sentiment.

Measures using the dictionary-based approach

In studies using the dictionary-based approach with the *GI* or custom word lists and proportional weighting, the most common measure is the percentage of the number of words in a given sentiment category to the total number of words in the text (e.g. Kothari et al. (2009), Ferguson et al. (2012), Chen et al. (2013)), or the standardized percentage (Z-score) (e.g. Tetlock et al. (2008)). Standardization is necessary if the raw frequency of matched words in the total number of words is not stationary, which can happen when regime changes occur over time in the distribution of words in the text (Tetlock et al. (2008)). This can happen, for example when the writing style changes with the author. The raw percentage of

words is always greater or equal to 0, but the standardized percentage can be either greater or less than 0. Positive/negative Z-scores indicate that the degree of optimism/pessimism or other affects in the text is above/below average. Another relative measure in contrast to the absolute measure (i.e. raw percentage) is the count of positive words minus the count of negative words, divided by the sum of positive and negative word counts or total world count (e.g. Rees and Twedt (2012)) . With the relative measure, one can more easily identify whether a piece of text is relatively positive (if greater than 0) or relatively negative (if less than 0), and the magnitude of sentiment relative to the length of the text. In studies using *DICTION*, sentiment measures or their constituent factors are also raw frequencies and standardized scores given by the program. For example, because there is no direct sentiment category of 'positive' or 'negative', Davis et al. (2011) calculate the optimistic measure as the frequencies of words of praise, satisfaction and inspiration, and the pessimistic measure as the frequencies of words of blame, hardship and denial.

Tetlock (2007) constructs his sentiment measure by employing principle component analysis (PCA) to extract the most important semantic component from the (77×77) variance-covariance matrix of the categories in the *GI* dictionary. By doing this, he extracts an underlying media factor as the linear combination of all *GI* categories which accounts for the maximum variation in the total variance of all the categories. In fact, the first factor is approximately equal to a linear combination of only 4 *GI* categories: *Negative, weak, fail* and *fall.* The advantage of PCA is that it does not decide *ex-ante* what type of sentiment to extract, but objectively extracts the measures of sentiment that account for the greatest variation in the total variance of all categories. This principle factor thus represents the overall style of the text. Doran et al. (2010) and Price et al. (2012) also use PCA to define sentiment measures. However, if one wants to focus on a certain type of sentiment, PCA would not be appropriate.

Measures using machine learning

Studies that perform machine learning construct the sentiment measure based on the classification of the texts. For example, Li (2010) defines the affect of a sentence to be 1 if the

learning algorithm predicts the sentence to be positive, 0 if neutral, and -1 if negative or uncertain. For each document, the author defines the overall affect as the average score of individual sentences. Das and Chen (2007) construct their daily sentiment series in a similar way. Sinha (2010) defines the sentiment score as the difference between the probability of the news article being positive and negative.

5 Econometric models and hypothesis testing

Having decided on the information source from which to construct the measure of sentiment (most commonly from corporate disclosures, news articles or internet messages), and having decided upon the content analysis method (usually some variant of the dictionary-based or machine learning approaches), the researcher constructs one or more measures of sentiment. Armed with a relatively new kind of data, the next step is to investigate any number of hypotheses about how the sentiment interacts with the financial variables and *vice versa*. The range of modeling approaches is very wide, and in this section we present a summary review of what has been investigated to date. We shall see that this research agenda is very new, and that a great deal of innovative approaches, models and hypotheses remain to be examined.

Linear regression models

The most common approach has been to employ the linear regression model on time series data comprising general market-level sentiment and stock index performance. Where cross-sectional sentiment and firm-level performance variables are available, some researchers have used panel regression methods to control for the individual heterogeneity of firms (e.g. Henry (2008), Demers and Vega (2011)). We can describe and summarize the general approach that has been adopted in single-equation studies within the sentiment-finance literature in Equation (1) below.

$$y_t = \alpha_0 + \sum_{j=1}^{J} \alpha_{1,j}\, y_{t-j} + \sum_{j=0}^{J} \sum_{f=1}^{F} \alpha_{2,j}^{f}\, X_{t-j}^{f} + \sum_{j=0}^{J} \sum_{a=1}^{A} \alpha_{3,j}^{a}\, S_{t-j}^{a} + \varepsilon_t \qquad (1)$$

In Equation (1), *y* is the dependent variable, $\boldsymbol{X}$ is a vector of $f = 1 \ldots F$ control variables, and $\boldsymbol{S}$ is a vector of various measures of affect or sentiment. Equation (1) states that firm-level or market-level performance in period *t* depends on a constant term, α_0; on $j = 1 \ldots J$ lagged dependent terms, $\sum_{j=1}^{J} \alpha_{1,j}\, y_{t-j}$; on $j = 0 \ldots J$ contemporaneous or lagged terms of $f = 1 \ldots F$ control variables in the $\boldsymbol{X}$ vector, $\sum_{j=0}^{J} \sum_{f=1}^{F} \alpha_{2,j}^{f}\, X_{t-j}^{f}$; on $j = 0 \ldots J$ contemporaneous or lagged terms of $a = 1 \ldots A$ affect or sentiment terms in the $\boldsymbol{S}$ vector, $\sum_{j=0}^{J} \sum_{a=1}^{A} \alpha_{3,j}^{a}\, S_{t-j}^{a}$; and on a residual unexplained component, ε_t. In practice, the number of lagged terms can differ for each term on the right hand side of the equation, but we have set them all to *J* here for notational convenience. The subscript $i = 1 \ldots N$ denoting the number of observations has also been omitted for notational convenience.

The dependent variable in the textual sentiment – finance literature is typically some type of firm-level or market-level performance measure such as future earnings (e.g. Li (2010), Demers and Vega (2011), Huang et al. (2011)), future earnings changes (e.g. Li (2006), Li (2010)), future returns on assets (e.g. Davis et al. (2011)) and future cash flows (e.g. Huang et al. (2011)). Tetlock et al. (2008) employ standardized unexpected earnings (SUE)[9] as the dependent variable. Other dependent variables have also been employed. For example, in Antweiler and Frank (2004), 15-minute interval stock returns, price volatilities, two measures of trading volumes[10] and a liquidity measure[11], are each employed in turn as dependent variables. Kothari et al. (2009) investigate the impact of textual sentiment on three dependent variables: firms' costs of capital, the standard deviations of stock returns, and the standard deviations of analysts' forecast errors. Tetlock (2007) uses negative sentiment to attempt to predict the Small-Minus-Big (SMB) factor, the idea being that because small stocks have the highest individual investor ownership, if the pessimism factor proxies for small investor sentiment, it should predict the returns on small stocks. Li (2010) employs a future liquidity measure[12] as the dependent variable. Loughran and McDonald (2013a) have used absolute value of revision (the percentage revision in the IPO offer price from the mid-point of the filing range) as the dependent variable.

Researchers using the event study methodology have also defined the dependent variable in equation (1) as the cumulative abnormal return (CAR) (or event period excess return) over some event window. Engelberg (2008), Feldman et al. (2008), Henry (2008), Henry and Leone (2009), Doran et al. (2010), Davis et al. (2011), Davis and Tama-Sweet (2011), Demers and Vega (2011), Huang et al. (2011), Loughran and McDonald (2011a, 2011b), Davis et al. (2012), Engelberg et al. (2012), Jegadeesh and Wu (2012), and Price et al. (2012) all employ the standard event study methodology to examine the extent to which sentiment in corporate disclosures (or news articles about disclosures) impacts on firms' cumulative abnormal returns around the 'event' or during a post-event period. The calculation of abnormal returns is critical in event studies. The abnormal return is the difference between the actual return on the security over the event window and the normal return on the firm. There are choices for modeling the latter. For example, Davis et al. (2011), Engelberg (2008), Feldman et al. (2008) and Demers and Vega (2011) follow Fama and French (1992) in constructing the book-to-market and size-sorted portfolios. Henry (2008) and Loughran and McDonald (2011a) choose the return of an equal-weighted or value-weighted market index as the normal return to calculate abnormal returns for each stock. Tetlock et al. (2008), Chen et al. (2013), Rees and Twedt (2012) and Huang et al. (2013) apply a method similar to event studies. Tetlock et al. (2008) and Chen et al. (2013) set the news stories date as the 'event', while Rees and Twedt (2012) and Huang et al. (2013) set the analyst reports publication date as the 'event'. They test whether the negative sentiment measure has real impact on the firm's next-day return or abnormal returns over a certain holding period.

The $\boldsymbol{X}$ vector of control variables has included firm characteristics and previous market variables such as the cash flow from operations, the book-to-market ratio, the market value of equity, accruals and leverage, current earnings surprises, analyst earnings forecast revisions and dispersions, recent and more distant stock prices, returns and volatilities, and stock market index returns and trading volumes. The idea is to test if the sentiment measure still has significant predictability on the dependent variable measuring firm-level or market-level

performance after controlling for the vector of fundamental variables and previous market activities. In studies that examine textual sentiment in analyst reports (Rees and Twedt (2012) and Huang et al. (2013)), the control variables typically include the deviation of analyst's earnings per share forecast, the deviation of analyst's stock recommendation, the revisions of recommendation, earnings forecast and target price, and the indicator variables for recommendation upgrades, reiterations and downgrades.

Some researchers have also considered and examined the possibility of reverse causation, whereby firm-level or market-level performance can impact on future measures of sentiment. Equation (2) captures these possibilities in general form.

$$s_t = \beta_0 + \sum_{j=1}^{J} \beta_{1,j}\, s_{t-j} + \sum_{j=0}^{J} \sum_{f=1}^{F} \beta_{2,j}^{f}\, X_{t-j}^{f} + \sum_{j=0}^{J} \sum_{p=1}^{P} \beta_{3,j}^{p}\, Y_{t-j}^{p} + \varphi_t \qquad (2)$$

Here, *s* is the dependent variable obtained as some measure of affect or sentiment from the ***S*** vector in Equation (1), and ***Y*** is a vector of firm-level or market-level performance variables from which the dependent variable in Equation (1) is obtained. Equation (2) states that sentiment in period *t* depends on a constant term, β_0; on $j = 1 \ldots J$ lagged dependent terms, $\sum_{j=1}^{J} \beta_{1,j}\, s_{t-j}$; on $j = 0 \ldots J$ contemporaneous or lagged terms of $f = 1 \ldots F$ control variables in the ***X*** vector, $\sum_{j=0}^{J} \sum_{f=1}^{F} \beta_{2,j}^{f}\, X_{t-j}^{f}$; on $j = 0 \ldots J$ contemporaneous or lagged terms of $p = 1 \ldots P$ firm-level or market-level performance variables, $\sum_{j=0}^{J} \sum_{p=1}^{P} \beta_{3,j}^{p}\, Y_{t-j}^{p}$. As in Equation (1), we have set the number of lagged terms to *J* and omitted the subscript $i = 1 \ldots N$ for notational convenience.

Because corporate documents contain much information about firm-level fundamentals, it is natural to hypothesize that firm fundamentals also have impact on the sentiment of corporate disclosures. Antweiler and Frank (2004), Li (2006), Li (2010), Davis and Tama-Sweet (2011), Davis et al. (2012), Garcia (2012), Jegadeesh and Wu (2012) apply versions of Equation (2)

to test whether firm characteristics in the ***X*** vector (e.g. book-to-market equity ratio, the market value of equity, debt to asset ratio, earnings scaled by book value of assets, sales, accruals, stock return volatility) or the firm-level or market-level performance variables in the ***Y*** vector have explanatory power on their sentiment measure and changes in this.

Vector autoregression models

The vector autoregression (VAR) model has also been used to capture the evolution and interdependencies between firm-level or market-level performance, control variables and sentiment. The VAR is a more advanced time-series model than single equation models insofar as all variables in the VAR are determined endogenously by their own history and by the history of all the other variables in the model. The standard *p*-th order VAR is written as Equation (3).

$$A(L)Z_t + BX_t = u_t \tag{3}$$

with

$A(L) = 1 - A_1 L - A_2 L^2 - \ldots\ldots - A_p L^p$,

$E(u_t) = 0, \quad E(u_t u'_t) = \Sigma, \quad E(u_t u'_s) = 0$, for $t \neq s$, $\quad E(x_t u_t) = 0$,

$Z_t = (Y_t, S_t)$,

and X_t being the vector of controls in Equations (1) and (2).

This is a standard VAR representation in which *Z* is a $(1 \times n)$ vector of endogenous variables, with *n* depending on how many measures of firm-level or market-level performance and semantic measures are included from the ***Y*** and ***S*** vectors in Equations (1) and (2). A is an $(n \times n)$ matrix of coefficients, u is an $(n \times 1)$ vector of white noise disturbance terms, and L denotes the lag operator (for example, $L^i x_t = x_{t-i}$). It is well established that this VAR model can be estimated by ordinary least squares, which yields consistent and asymptotically

efficient estimates of the A matrix because the right-hand-side variables are predetermined and are the same in each equation of the model.

Tetlock (2007) sets up a VAR model with 5 lags to investigate the interdependencies between negative sentiment, the *Dow Jones Industrial Average* (*DJIA*) returns, and the NEW York Stock Exchange (*NYSE*) trading volumes. By employing the VAR system, the author attempts to predict the *DJIA* returns using negative sentiment, to predict negative sentiment using the *DJIA* returns, and to predict *NYSE* volumes using negative sentiment. He finds that negative sentiment causes *DJIA* returns, which have feedback effects on negative sentiment, and that negative sentiment also causes trading volumes. There remains a great deal of research to be done on extending the range of variables included in VAR models, and in investigating more generally how the many variables within the ***Y***, ***S*** and ***X*** vectors interact.

Logitstic and probit regressions

The logistic (logit) or probit regressions have been used in several studies to examine if textual sentiment helps predict or identify whether a specific event is likely to happen. These regressions do not assume a linear relationship between the dependent and independent variables, and the dependent variable must be a dichotomy (2 categories). The regressions are usually in the following form:

$$B_i = \gamma_0 + \gamma_1 s_i + \sum_{f=1}^{F} \gamma_{2,f} X_{f,i} \qquad (4)$$

Here, the dependent variable *B* is a binary variable, whose value is equal to 1 if a certain event has happened, and 0 otherwise. The textual sentiment measure *s* is the degree of sentiment contained in one or more document(s) of firm $i = 1 \ldots N$. ***X*** is a vector of $f = 1 \ldots F$ other independent variables, including firm characteristics and dummy variables. The idea is to examine if textual sentiment is significantly associated with a certain event, after controlling

for other relevant variables. Unlike linear regressions, the logit and probit regressions are estimated by the maximum likelihood method.

Huang et al. (2011), Loughran and McDonald (2011a, 2011b, 2013a), Rogers et al. (2011), Buehlmaier (2013) have all employed equation (4) in their research. For example, Loughran and McDonald (2011a) have employed two groups of logit regressions to examine if companies accused of accounting improprieties or firms that self-report material weaknesses in internal controls use different language from other firms in their 10-Ks. For the first group, the binary dependent variable is equal to 1 if the firm was accused of accounting fraud in the year after the 10-K filing date; for the second group, the dependent variable is equal to 1 if within 18 months of the 10-K file date a disclosure of material weakness is reported. They find that word lists (affects) can play a role in identifying firms experiencing unusual events. Loughran and McDonald (2013a) have examined whether an overabundance of negative sentiment in the S-1 filing would lead to the withdrawal of an IPO. In the logit regressions, the dependent variable is set to 1 if the IPO has filed a Form RW (registration withdrawal request) and did not subsequently issue a Form 424.They find that there is no material difference in the S-1 texts between the withdrawn IPOs and completed IPOs.

Volatility models

Antweiler and Frank (2004) employ volatility models to test the effects of sentiment or sentiment-related variables on the second moment of stock returns. They employ both GARCH-type models and the 'realized volatility' approach, but only report results of the latter, as both results are consistent with each other. In their GARCH-type approach, they let the sentiment measure 'bullishness' enter the mean equation first, and they insert the 'agreement' measure, amongst other exogenous variables, into the volatility equation, making the model into GARCH-X, GJR-GARCH-X and EGARCH-X forms[13]. In the 'realized volatility' approach, the authors investigate the extent to which the internet postings have an impact on the realized volatility (proxied by the log of the standard deviation of intra-day returns) by incorporating the 'agreement' measure, the number of messages and the number

of trades as explanatory variables. They find that the number of messages is a predictive factor of volatility, but agreement among the posted messages is not significant.

Textual sentiment based trading strategies

Another intuitive way to examine the extent to which positive and negative verbal information leads to significant differences in risk-adjusted stock returns is to design trading strategies based on textual sentiment. This methodology can be considered complementary to econometric modeling to examine the extent to which textual sentiment has financial market impacts. Tetlock et al. (2008) label all news stories with a fraction of negative words (Neg) in the previous year's top (bottom) quartile as negative (positive) stories. They consider a simple strategy that longs all firms with positive news stories on the prior trading day and shorts all firms with negative news stories. They model holding both the long and short portfolios for one full trading day and rebalancing at the end of the next trading day. They use the Fama-French three-factor (1993) and the Carhart four-factor (1997) models to adjust the trading strategy returns for the returns on the market, size, book-to-market, and momentum factors. Engelberg (2008) form a similar trading strategy which longs firms with articles that have no negative words and shorts firms with articles that have above 5% negative words. This author models holding the portfolio for 80 days after the earnings announcement, and examined portfolio returns for the first and second 40 days separately. The portfolio returns are risk-adjusted by the Carhart four-factor (1997) model. Li (2006), Feldman et al. (2008), Sinha (2010), Demers and Vega (2011), Loughran and McDonald (2011a), Engelberg et al. (2012), Ferguson et al. (2012), and Rees and Twedt (2012) have also constructed similar portfolios and examined their abnormal returns. Their results generally show that the textual-sentiment based strategies can produce significant abnormal returns. A strategy based on short selling and negative news by Engelberg et al. (2012) would have earned an astonishing 180% during their 2.5-year sample period. However, Loughran and McDonald (2011a) and Rees and Twedt (2012) do not find evidence of return predictability based on the trading strategies. Nevertheless, it appears from this analysis that further work on the

relationship between textual sentiment and financial market activity might be useful to market participants as well as researchers.

6 Main findings to date

Table 3 summarizes the information source, time frame, content analysis method, models and main findings of the most significant and influential studies. Since space prevents a complete description of each paper, we instead point to the main topics and findings.

Engelberg (2008), Feldman et al. (2008), Henry (2008), Henry and Leone (2009), Doran et al. (2010), Davis et al. (2011), Davis and Tama-Sweet (2011), Demers and Vega (2011), Huang et al. (2011), Loughran and McDonald (2011a, 2011b), Davis et al. (2012), Jegadeesh and Wu (2012), Price et al. (2012), and Rees and Twedt (2012) all contribute to the analysis of announcement effects on asset prices, as they apply the standard event-study method to study the impact of textual sentiment in earnings announcement or financial reports on asset prices.

Li (2006), Tetlock (2007), Tetlock et el. (2008), Li (2010), Sinha (2010), Davis et al. (2011), Demers and Vega (2011), Huang et al. (2011), Ferris et al. (2012), and Chen et al. (2013) have all studied the connection between textual sentiment and future firm fundamentals or stock performance, contributing to the study of information and market efficiency. Tetlock (2007), Tetlock et al. (2008), Ferguson et al. (2012), and Garcia (2012) concentrate on the interrelations between news stories sentiment and immediate market response, adding knowledge about the role of media in financial markets. Garcia (2012) tries to compare the different market reactions during recession and non-recession periods.

Antweiler and Frank (2004), Das and Chen (2007) and Chen et al. (2013) concentrate on the connection between internet board posting activities and the stock market. They study not only sentiment in these messages, but also the degree of ‘agreement’ of opinions, ‘article attention’, the number of messages, and the number of words. Rees and Twedt (2012) and

Huang et al. (2013) extract textual sentiment from financial analyst reports and examine whether sentiment in analyst reports has incremental information content. The latest research by Ferris et al. (2012) and Loughran and McDonald (2013a) have focused on the effects of textual sentiment on IPO pricing and subsequent performance. They use IPO prospectus and S-1 filing texts as the information source, respectively. Davis and Tama-Sweet (2011), Huang et al. (2011), Rogers et al. (2011), and Larcker and Zakolyukina (2012) have made novel contributions to the study of managers' strategic reporting, tone management or misreporting.

Henry and Leone (2009), Loughran and McDonald (2011a) and Jegadeesh and Wu (2012) endeavor to improve the accuracy of the dictionary-based approach to content analysis. The first two studies construct word lists that are appropriate in the finance domain, while the second and third studies question the use of the proportional term-weighting scheme that has been extensively used in earlier studies.

Sentiment and security prices and returns

Textual sentiment or the tone of qualitative information has been found to have important effects on stock prices and returns. Both the media-expressed and the internet-expressed sentiment literatures have found strong evidence of the *immediate* effects of sentiment. Particularly, negative sentiment has proved to be the strongest influence. Negative sentiment or a large increase in negative sentiment causes downward pressure on market prices immediately. Tetlock (2007), Tetlock et al. (2008) and Garcia (2012) have all concluded this. Chen et al. (2013) find that the negative sentiment in internet articles is negatively associated with both contemporaneous and next-day abnormal returns, but Antweiler and Frank (2004) discover that a positive shock to message board postings predicts negative returns on the next day. The corporation-expressed sentiment literature (e.g. Engelberg (2008), Doran et al. (2010), Davis et al. (2011), Demers and Vega (2011), Durnev and Mangen (2011), Jegadeesh and Wu (2012), Price et al. (2012)) concludes that the tone of corporate disclosures or changes in the tone from the recent past are significantly correlated with short window contemporaneous returns around the date that the disclosures are made, or drift excess

returns[14], even after controlling for firm financial information and earnings surprises. Abnormal market returns are higher as the tone of the press release becomes more positive. The latest research by Loughran and McDonald (2011a) shows that large amounts of uncertain text in an S-1 filing generally lead to higher IPO first-day returns.

Although negative sentiment seems to have the strongest influence, Jegadeesh and Wu (2012) claim that both their measures of positive and negative sentiment are associated with the 10-K filing date returns. Using a *net optimism* measure (the difference between the percentage of optimistic words and the percentage of pessimistic words), both Davis et al. (2011) and Demers and Vega (2011) have found that sentiment in the earning press releases significantly affects the announcement and/or post-announcement period returns. Ferguson et al. (2012) conclude that high-attention news (both positive and negative) affects subsequent trading period returns.

Considering the possibility of reverse causation, do stock prices or returns predict textual sentiment? Tetlock (2007) discovers that negative returns predict more pessimism in the next day's WSJ column. Garcia (2012) finds that stock returns are indeed important predictors of media sentiment, both the positive and the negative measures, and that as expected, positive market returns tend to increase positivity and decrease negativity in future news stories. However, Das and Chen (2007)'s finding is that the value of the stock index on a given day is not related to the sentiment level in internet messages on the next day.

Sentiment and other market variables

Studies have also found that textual sentiment has significant impacts on trading volumes. Tetlock (2007) finds that unusually high or low values of pessimism lead to temporarily high market trading volume. Antweiler and Frank (2004) discover that internet board postings can help predict trading volumes, and using daily data showed that the effect from message posting activities to trading volumes is more significant than the reverse direction. Das and Chen (2007) conclude that there is a strong contemporaneous relationship between sentiment

and trading volume. Although these authors examined internet-posting sentiment, they conclude that sentiment has no predictability on future trading volumes.

Some studies have found the evidence of the connection between textual sentiment and stock market volatility. Antweiler and Frank (2004) find that internet message postings help predict volatility. Tetlock (2007) discovers that the conditional volatility of the Dow Jones appears to be higher when the pessimism factor is high. Kothari et al. (2009) find that both positive and negative news disclosures in the media have impact on firms' return volatility and that when management offers negative news disclosures, return volatility rises. Loughran and McDonald (2011b) also find that the use of the 13 phrases that appear on Barron's list of worrisome words is significantly related to subsequent return volatility. The findings of the latest research by Loughran and McDonald (2011a) show that large amounts of uncertain text in an S-1 filing generally lead to higher subsequent return volatility (60-day period after the IPO).

Sentiment, firm fundamentals and market efficiency

The literature to date shows that textual sentiment is correlated with future firm characteristics and performance. Li (2006) finds that an increase in the risk sentiment of annual reports is associated with significantly lower future earnings. Tetlock et al. (2008) show that negative words in news stories forecast low subsequent firm earnings. Li (2010) discovers that the average tone of the forward-looking statements in a firm's MD&A is positively correlated with its future earnings and has explanatory power incremental to other variables in predicting future performance. Davis and Tama-Sweet (2011) find that higher levels of pessimistic language in the MD&A are associated with lower future return on assets. Demers and Vega (2011) conclude that the net optimism and certainty expressed in managerial earnings announcements are associated with various proxies for future earnings and expected earnings uncertainty. Similarly, Huang et al. (2011) find that abnormal positive tone in earnings press releases is associated with poor future earnings. The results of Ferris et al. (2012) show that conservatism in IPO prospectus for non-technology IPOs contains useful information about the firm's future operating performance. Chen et al. (2013) find that negative sentiment 30

days to 3 days prior to earnings announcements is inversely associated with the earnings surprise. Both Rees and Twedt (2012) and Huang et al. (2013) have concluded that textual sentiment in analyst reports is informative about firms' future performance beyond quantitative measures. Meanwhile, the findings of Li (2006), Li (2010), Davis and Tama-Sweet (2011), Davis et al. (2012), and Jegadeesh and Wu (2012) suggest that firm fundamentals such as market value of equity, the book-to-market ratio, accruals, quarterly earnings scaled by the book value of assets, stock return volatility, debt to assets, and earnings surprise measures[15] are among the determinants of sentiment in corporate disclosures.

Does the linguistic style of corporate disclosures, news articles and internet messages contain useful incremental information over quantitative financial information is assessing market efficiency? Aggregating the results in the previous paragraph, we know that the answer is generally yes, although it must be noted that 'linguistic style' is not solely represented by sentiment. Tetlock et al. (2008) argue that the stock market is relatively efficient with respect to firms' hard-to-quantify fundamentals, although Tetlock (2007) finds little support for the hypothesis that textual sentiment represents additional fundamental information. Li (2006) concludes that the stock market does not fully reflect the information contained in the texts of annual reports about future profitability, and Sinha (2010) argues that the market underreacts to the tone of news articles. By and large, it has been clear that sentiment in texts does convey incremental information over quantitative financial information, and that it might have power in predicting market movements. Sentiment from qualitative information is mostly publicly available, or possibly partly private in some cases. In either case, according to the strongest form of the EMH, a good asset pricing model might well incorporate textual sentiment as a pricing factor in addition to risk premia and other firm-level characteristics.

7 Summary and conclusions

We have reviewed the textual sentiment literature in finance by focusing on three main aspects – the information sources, the content analysis methods, and the financial models that

have been used to examine whether and how textual sentiment impacts on people, institution and markets. The surveyed literature has made contributions to different areas of both behavioral and traditional finance. It has contributed to the analysis of announcement effects; to our understanding of the complementarity relation between qualitative and quantitative information regarding firm performance; it has improved the linguistic analysis methods in order to get more accurate and efficient sentiment measures; and has facilitated the study of market efficiency from a new perspective in contrast to the earlier work that used easily quantifiable aspects of news as a proxy for the news itself, such as the timing of news, the volume of news (the number of words), and the type of news (periodic announcements and general publicly available news).

What is agreed and what remains controversial? It is generally agreed that textual sentiment has potentially strong impacts on stock returns and trading volumes. The media-expressed sentiment literature demonstrates that textual sentiment has contemporaneous or short-term effects on stock prices, returns, abnormal returns, and trading volumes (e.g. Antweiler and Frank (2004), Tetlock (2007), Tetlock et al. (2008), Ferguson et al. (2012), Garcia (2012), Jegadeesh and Wu (2012), Chen et al. (2013)). The exception is Das and Chen (2007), who conclude that sentiment has no predictability on future trading volumes. Event studies in the corporation-expressed sentiment literature tend to exhibit significant market responses to sentiment in corporate disclosures, in a short window (1 day or 3 days) around the event date (e.g. Davis et al. (2011), Engelberg (2008), Feldman et al. (2008), Demers and Vega (2011), Jegadeesh and Wu (2012), Price et al. (2012)).

In the dictionary-based approach literature, particularly the studies which use the GI dictionary or L&M word lists, negative words are found to be the most powerful sentiment category in explaining market activities. Tetlock (2007) emphasizes that negative words summarize common variation in the entire set of GI word categories better than any other single category including positive words. Tetlock et al. (2008) find that using positive words as the sentiment measure produces much weaker results than using negative words. However,

the more recent study by Jegadeesh and Wu (2012) has pointed out that by using a new term-weighting scheme, a significant relation between document tone and market reaction is discovered for positive words as well, so they emphasize the importance of choosing the appropriate term-weighting scheme.

Stock returns are found to predict textual sentiment by Tetlock (2007) and Garcia (2012), who use news stories as the information source. Yet Das and Chen (2007) reveal that the value of the stock index on a given day is not related to the sentiment level in internet messages on the next day. The work by Li (2006), Tetlock et al. (2008), Feldman et al. (2008), Li (2010), Davis and Tama-Sweet (2011), Demers and Vega (2011), Huang et al. (2011), Ferris et al. (2012), Garcia (2012), Price et al. (2012), Rees and Twedt (2012), Chen et al. (2013), and Huang et al. (2013) consistently shows that textual sentiment contains new value-relevant information. The exception is Tetlock (2007), who rejected the hypothesis that media content contains new information about fundamental asset values.

We conclude by suggesting the following directions for future research are suggested. *First*, the content analysis process can be further improved. Although the L&M word lists have been increasingly popular in the latest research, for the dictionary-based approach, the construction and availability of more authoritative and extensive field-specific dictionaries is desirable for future studies. There is also scope for the development of improved term-weighting schemes. Other content analysis approaches can also be developed and refined by linguists, psychologists and computer scientists. The purpose of improving the content analysis procedures is to obtain more accurate and efficient sentiment measures, and to examine whether other sentiment categories can be as powerful as negative words.

Second, there are still some qualitative information sources that have not yet been widely studied, such as business and political speeches, blogs, television news, videos and various social media. Speeches from the most influential individuals in the economics and finance fields are particularly worth studying. *Third*, volatility models have rarely been used. For

example, the GARCH-X, GARCH-M-X models can be applied to examine the direct and indirect relationships between sentiment and volatilities. They may also assist in discovering the mechanism through which sentiment in texts gets transmitted to asset prices and returns. *Fourth*, the extant textual sentiment studies tend to focus on the stock market. The same methodologies can be applied to other markets such bonds, commodities and derivatives. Meanwhile, it is interesting to study global financial markets other than the U.S., especially emerging markets, and to analyze qualitative information in other languages, as different markets may display differing cultural and other behavioral patterns. *Fifth*, with the further development of computer programs to source continuous flows of news and to analyze sentiment automatically, it will be increasingly possible to examine the effect of sentiment in real time. This is exciting because one can view and analyze the price movement and trading volume immediately following a piece of extremely positive or negative news. *Finally*, the complex and most probably time-varying relationship between textual sentiment and investor behavior remains an important area of future study that promises to add many more insights into the theory, evidence and practice of behavioral finance.

Figure 1: sentiment extraction (dictionary-based approach)

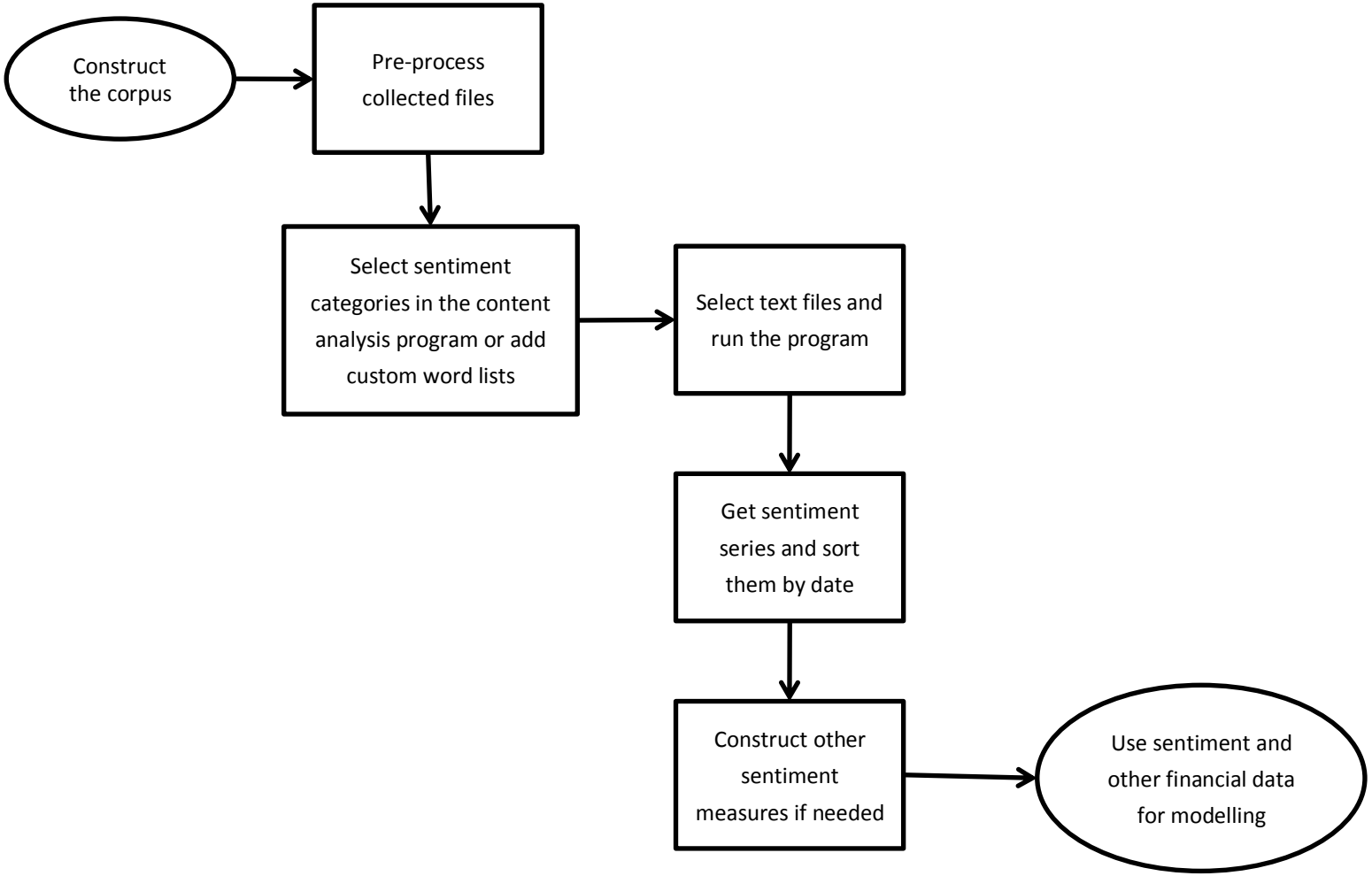

Table 1: Qualitative information sources

Notes: This table summarizes the information sources used in the literature. They are corporate disclosures, news stories and analyst reports, and internet board postings.

Corporation-expressed Sentiment			**Media-expressed Sentiment**		**Internet-expressed Sentiment**	**Mixed sources**
Annual Reports/10-Ks/ 10-Qs	*Earnings press Releases/ Earnings conference calls*	*Other/Mixed*	*News stories and commentaries*	*Analyst reports*	*(Board postings)*	
Li (2006)	Henry (2006)	Durnev and Mangen (2011)	Tetlock (2007)	Tetlock (2007)	Antweiler and Frank (2004)	Kothari et al. (2009)
Feldman et al (2008)	Henry(2008)	Rogers et al. (2011)	Tetlock et al. (2008)	Engelberg(2008)	Das and Chen (2007)	Ozik and Sadka (2012)
Li (2010)	Henry and Leone (2009)	Ferris et al. (2012)	Engelberg (2008)	Tetlock et al. (2008)	Chen et al. (2013)	
Loughran and McDonald (2011a)	Doran et al. (2010)	Loughran and McDonald (2013a)	Sinha (2010)	Sinha (2010)		
Loughran and McDonald (2011b)	Davis et al. (2011)		Garcia (2012)	Carretta et al. (2011)		
Jegadeesh and Wu (2012)	Davis and Tama-Sweet (2011)		Rees and Twedt (2012)	Engelberg et al. (2012)		
	Demers and Vega (2011)		Huang et al. (2013)	Ferguson et al. (2012)		
	Huang et al. (2011)			Garcia (2012)		
	Davis et al. (2012)			Buehlmaier (2013)		
	Larcker and Zakolyukina (2012)			Liu and McConnell (2013)		
	Price et al. (2012)					

Table 2: Content analysis methods

Notes: This table summarizes the content analysis methods used in the literature. For the dictionary-based approach, the programs or dictionaries used include DICTION, General Inquirer (GI) / *Harvard IV-4* dictionaries, finance-specific word lists and other specific words.

Dictionary-based Approach				**Machine Learning**
DICTION	*GI/Harvard*	*Finance-specific words*	*Other*	
Henry and Leone (2009)	Tetlock (2007)	Henry (2006)	Li (2006)	Antweiler and Frank (2004)
Davis et al. (2011)	Engelberg(2008)	Henry(2008)	Loughran and McDonald (2011b)	Das and Chen (2007)
Davis and Tama-Sweet (2011)	Feldman et al (2008)	Henry and Leone (2009)	Larcker and Zakolyukina (2012)	Li (2010)
Demers and Vega (2011)	Tetlock et al. (2008)	Doran et al. (2010)	Buehlmaier (2013)	Sinha (2010)
Durnev and Mangen (2011)	Henry and Leone (2009)	Demers and Vega (2011)		Huang et al. (2013)
Rogers et al. (2011)	Kothari et al. (2009)	Huang et al. (2011)		
Davis et al. (2012)	Doran et al. (2010)	Loughran and McDonald (2011a)		
Ferris et al. (2012)	Carretta et al. (2011)	Rogers et al. (2011)		
	Demers and Vega (2011)	Davis et al. (2012)		
	Loughran and McDonald (2011a)	Engelberg et al. (2012)		
	Engelberg et al. (2012)	Ferguson et al. (2012)		
	Ferris et al. (2012)	Ferris et al. (2012)		
	Ozik and Sadka (2012)	Garcia (2012)		
	Price et al. (2012)	Jegadeesh and Wu (2012)		
	Rees and Twedt (2012)	Ozik and Sadka (2012)		
		Price et al. (2012)		
		Chen et al. (2013)		
		Liu and McConnell (2013)		

Table 3: Summary of each study

Notes: This table summarizes the information sources, content analysis methods and models employed in the literature, as well as their studying period and the key findings and insights. In the fourth column 'Content analysis methods', the specific programs, dictionaries, and word lists used for the dictionary-based approach are also listed. Among them, *GI/Harvard* indicates General Inquirer / *Harvard IV-4* dictionaries, *DICTION* is the textual analysis program and its self-contained dictionaries, *L&M* indicates Loughran and McDonald (2011a) finance dictionaries, *Henry* indicates the finance word lists constructed by Henry (2006, 2008). The findings/insights listed are those relevant to the role of textual sentiment.

Research	Information sources	Time frame	Content analysis methods	Models	Key findings
Antweiler and Frank (2004)	Internet messages	2000	Machine learning	Linear regressions Volatility models	A positive shock to message board postings predicts negative returns the next day; contemporaneous regressions show that disagreement induces trading; message posting helps predict volatility; stock messages reflect public information very rapidly.
Henry (2006)	Earnings press releases	2002	Dictioary-based *Custom words*	Data-mining algorithm	The inclusion of verbal variables (including 'tone') results in greater predictive accuracy beyond inclusion of financial information.
Li (2006)	Annual reports	1994-2005	Dictioary-based *Specific words (risk and uncertain)*	Linear regressions Sentiment-based trading strategy	An increase in risk sentiment of annual reports is associated with significantly lower future earnings; the change in risk sentiment is also negatively related to cross-sectional future returns; the stock market does not fully reflect information contained in the texts of annual reports about future profitability.
Das and Chen (2007)	Internet messages	Jul. - Aug. 2001	Machine learning	Linear regressions	Sentiment aggregated across stocks tracks index returns; aggregation of sentiment reduces some of the noise from individual stock postings; market activity is related to small investor sentiment and message board activity.
Tetlock (2007)	News articles	1984-1999	Dictioary-based *GI/Harvard*	VAR	High values of media pessimism induce downward pressure on market prices; unusually high or low values of pessimism lead to temporarily high market trading volume; the changes in market returns that follow pessimistic media content are dispersed throughout the trading day; pessimism weakly predicts increases in market volatility.
Engelberg(2008)	News articles about firms' earnings announcement	1999-2005	Dictioary-based *GI/Harvard*	Linear regressions, including event-study method Sentiment-based trading strategy	Qualitative earnings information embedded in the news stories about firms' earnings announcement has additional predictability for asset prices beyond the quantitative information; qualitative information about positive fundamentals and future performance is most important for the prediction of future returns.

Table 3 (continued)

Research	Information sources	Time frame	Content analysis methods	Models	Key findings
Feldman et al (2008)	10-Ks and 10-Qs	1995-2006	Dictionary-based *GI/Harvard*	Linear regressions, event-study methods, Sentiment-based trading strategy	Changes in the tone of the MD&A section from the recent past are significantly correlated with short window contemporaneous returns around SEC filing dates; the MD&A sections do have information content.
Henry(2008)	Earnings press releases	1998-2002	Dictioary-based *Custom words*	Linear regressions, including event-study method	Abnormal market returns are higher as the tone of the press release becomes more positive.
Tetlock et al. (2008)	News articles	1980-2004	Dictioary-based *GI/Harvard*	Linear regressions Sentiment-based trading strategy	Negative words convey negative information about firm earnings above and beyond stock analysts' forecasts and historical accounting data; stock market prices respond to the information embedded in negative words with a small, one-day delay; negative words in stories about fundamentals predict earnings and returns more effectively than negative words in other stories.
Henry and Leone (2009)	Earnings press releases	2004-2006	Dictioary-based *Henry DICTION GI/Harvard*	Linear regressions, including event-study method	Capital markets researchers aiming to measure the qualitative information in financial disclosure can significantly increase the power for their tests by using a domain-specific wordlist; the economic impact of positive tone is smaller compared to negative tone.
Kothari et al. (2009)	10-Ks and 10-Qs, news articles, analyst reports	1996-2001	Dictioary-based *GI/Harvard*	Linear regressions	When management offers negative news disclosures, return volatility rises, and analyst forecast dispersion widens; both positive and negative news disclosures from analysts appear to be heavily discounted by the market; both positive and negative news disclosures in the business press impact the cost of capital, return volatility, and analyst forecast dispersion.
Doran et al. (2010)	Earnings conference calls	2004-2007	Dictioary-based *GI/Harvard L&M Henry*	Linear regressions, including event-study method	The overall tone of REIT earnings conference calls as a whole has significant explanatory power for the accompanying abnormal returns; a positive conference call tone nearly offsets the damaging effects of a negative earnings surprise; results are significantly stronger for measures which utilize the customized dictionary (Henry (2008) and Loughran and McDonald (2011)).
Li (2010)	10-Ks and 10-Qs	1994-2007	Machine learning Dictioary-based (supplementary)	Linear regressions	Average tone of the forward-looking statements in a firm's MD&A is positively correlated with its future earnings and liquidity and has explanatory power incremental to other variables in predicting future performance; the tone of MD&As is related to the cross-sectional association of accruals with future stock returns; empirical results based on the dictionary-based approach do not support the hypothesis that MD&As contain information content about future performance.

Table 3 (continued)

Research	Information sources	Time frame	Content analysis methods	Models	Key findings
Sinha (2010)	News articles	2003-2010	Machine learning	Linear regressions Sentiment-based trading strategy	The market underreacts to the tone of news articles; the portfolio resulting from the long-short sentiment-based trading strategy is highly correlated with the momentum factor; short-term reversal does not occur when returns are accompanied by information that matches the direction of returns.
Carretta et al. (2011)	News articles	2004-2010	Dictioary-based *GI/Harvard*	Structural VAR	A positive (negative) media sentiment in news spread before spin-off deals is associated with positive (negative) short term returns; an increase in investor attention determines an increase of trading volumes and volatility of spin-off firms in both the short and the long run.
Davis et al. (2011)	Earnings press releases	1998-2003	Dictioary-based *DICTION*	Linear regressions, including event-study method	Managers' earnings press release language communicates credible information about expected future firm performance to the market and that the market responds to this information.
Davis and Tama-Sweet (2011)	Earnings press releases and corresponding MD&As	1998-2003	Dictioary-based *DICTION*	Linear regressions, including event-study method	Managers omit or shift pessimistic language from their earnings press releases when they have strong incentives to report strategically; alternative disclosure outlets provide managers with an opportunity for strategic reporting; higher levels of pessimistic language in the MD&A are associated with lower future return on assets
Demers and Vega (2011)	Earnings press releases	1998-2006	Dictioary-based *GI/Harvard* *DICTION* *L&M*	Linear regressions, including event-study method Sentiment-based trading strategy	The net optimism and certainty expressed in managerial earnings announcements are associated with various proxies for future earnings and expected earnings uncertainty; there is an inverse association between the certainty in management's diction and the idiosyncratic volatility in the company's share price during the announcement window; managerial language tone conveys information regarding the firm's valuation fundamentals.
Durnev and Mangen (2011)	Restatements of financial reports	1997-2005	Dictioary-based *DICTION*	Develop theoretical models Linear regressions, including event-study method	When restatement tone becomes more pessimistic, both restating firms and their competitors experience lower abnormal returns at restatement announcements; the tone of restatement texts decreases information asymmetry about restating firms and their competitors; restatement tone provides a signal about restating firms' private information regarding unknown investment payoffs; disclosure tone carries new information that is relevant for investment decisions.
Huang et al. (2011)	Earnings press releases and corresponding MD&As	1997-2007	Dictioary-based *L&M*	Linear regressions, including event-study method Logit regressions	Abnormal positive tone in the earnings press release is associated with poor future earnings and operating cash flows in each of one-year to three-year ahead periods. Managers tend to use disclosure tone to complement quantitative earnings management. Tone manipulation succeeds in misleading investors.

Table 3 (continued)

Research	Information sources	Time frame	Content analysis methods	Models	Key findings
Loughran and McDonald (2011a)	10-Ks	1994-2008	Dictioary-based (include new term weighting scheme) *Custom dictionary GI/Harvard*	Linear regressions, including event-study method Logit regressions Sentiment-based trading strategy	Almost three-fourths of negative word counts in 10-Ks based on the Harvard dictionary are typically not negative in a financial context; the authors create a list of words that typically have a negative meaning in financial reports, and create a term weighting scheme that can lower the noise introduced by word misclassifications; they suggest the use of their financial word lists to avoid those words in the Harvard list that might proxy for industry or other unintended effects.
Loughran and McDonald (2011b)	10-Ks	1994-2008	Dictioary-based *13 suspicious phrases*	Linear regressions, including event-study method Logit regressions	The use of the 13 phrases that appear on Barron's list of worrisome words is significantly related to excess filing period returns, analyst earnings forecast dispersion, subsequent return volatility, and fraud allegations.
Rogers et al. (2011)	A variety of corporate disclosures	Not applicable	Dictioary-based *DICTION Henry L&M*	Logit regressions	Sued firms use substantially more optimistic language in their earnings announcements than non-sued firms do; the interaction between optimism and abnormal insider selling is associated with an increased probability of being sued; managers can reduce litigation risk by dampening the tone of disclosure.
Davis et al. (2012)	Earnings conference calls	2002-2009	Dictioary-based *DICTION Henry L&M*	Linear regressions, including event-study method	Observable manager-specific characteristics (e.g., gender, age, educational and career experiences) explain variation in residual tone; managers' individual styles potentially impact the market reaction to earnings announcements.
Engelberg et al. (2012)	News articles	2005-2007	Dictioary-based *GI/Harvard L&M*	Linear regressions, including event-study method Sentiment-based trading strategy	The well-documented negative relation between short sales and future returns is twice as large on news days and four times as large on days with negative news; a strategy based on short selling and negative news would have earned an astonishing 180% during the authors' 2.5-year sample period.
Ferguson et al. (2012)	News articles	1981-2010	Dictioary-based *L&M*	Linear regressions Sentiment-based trading strategy	Positive as well as negative words in news media content display relevant information; predictive relationship between media content and firms' returns is significant for low visibility firms with low market capitalization and high book-to-market ratio; high-attention news (both positive and negative) affects subsequent trading period returns
Ferris et al. (2012)	IPO prospectus	1999-2005	Dictioary-based *L&M DICTION GI/Harvard*	Linear regressions	Greater conservatism in the prospectus is related to increased underpricing; prospectus conservatism for non-technology IPOs contains useful information about the firm's future operating performance

Table 3 (continued)

Research	Information sources	Time frame	Content analysis methods	Models	Key findings
Garcia (2012)	News articles	1905-2005	Dictioary-based *L&M*	Linear and non-linear regressions	News content helps predict stock returns at the daily frequency, particularly during recessions; this asymmetric predictability is not driven by differences in reporting along the business cycle, and the effect is especially strong on Mondays and on days after holidays; investor sentiment has a prominent effect during bad times.
Jegadeesh and Wu (2012)	10-Ks	1995-2010	Dictioary-based (include new term weighting scheme) L&M	Linear regressions, including event-study method	The authors' measure of document tone based on their new return-based term weighting scheme is reliably related to 10-K filing date returns for both positive and negative words; their approach can be extended to minimize the level of subjectivity required for content analysis; with their term weighting method, useful information can be extracted even if the underlying wordlists contain extraneous words or when they are incomplete; the market does not fully respond to the tone of 10-Ks during the filing period.
Larcker and Zakolyukina (2012)	Earnings conference calls	2003-2007	Dictioary-based *LIWC*	Linear regressions	The linguistic features of CEOs and CFOs in conference call narratives can be used to identify financial misreporting; deceptive CEOs use significantly more extremely positive emotion words and fewer anxiety words; deceptive CFOs use significantly more words of negation and extremely negative emotion words.
Ozik and Sadka (2012)	A variety of news articles and corporate disclosures	1999-2008	Dictioary-based *GI/Harvard L&M*	Linear regressions	On average, when a fund is covered either simultaneously by multiple sources (i.e. General newspapers, Specialized magazines and Corporate communications) or exclusively by one of the three sources, the sentiment of Corporate news is most positive and that of General news is least positive, reflecting a reporting style bias and an editorial selection bias.
Price et al. (2012)	Earnings conference calls	2004-2007	Dictioary-based *GI/Harvard Henry*	Linear regressions, including event-study method	Earnings-specific dictionary is much more powerful in detecting relevant conference call tone; conference call discussion tone has highly significant explanatory power for initial reaction window abnormal returns as well as the post-earnings-announcement drift; conference call 'question and answer' tone matters more when firms do not pay dividends.
Rees and Twedt (2012)	Analyst reports	2006	Dictioary-based *GI/Harvard*	Linear regressions, including event-study method Sentiment-based trading strategy	Analyst report complexity and tone provide incremental information content to the market beyond quantitative summary measures; there is no evidence that abnormal returns can be earned through long-term trading strategies based on these qualitative report attributes.

Table 3 (continued)

Research	Information sources	Time frame	Content analysis methods	Models	Key findings
Buehlmaier (2013)	News articles	2000-2006	Machine learning	Develop theoretical models Linear regressions Logit regressions	Positive media content about the acquirer predicts takeover success
Chen et al. (2013)	Internet messages	2005-2011	Dictioary-based *L&M*	Linear regressions, including event-study method	The fraction of negative words contained in the articles and comments on Seeking Alpha (SA) website negatively predict subsequent stock returns; articles and comments, as a system, predict future stock returns much more strongly than articles alone; the fraction of negative words in articles and comments prior to the earnings announcement strongly predict subsequent scaled earnings surprises.
Huang et al. (2013)	Analyst reports	1995-2008	Machine learning	Linear regressions, including event-study method	The classification accuracy achieved using the naive bayes machine learning approach is substantially higher than that achieved using the dictionary-based content analysis approaches; investors place much more weight on negative than positive statements; report text is informative about a firm's short- and long-term performance; negative statements are more informative than positive ones about a firm's future performance.
Liu and McConnell (2013)	News articles	1990-2010	Dictioary-based *L&M*	Probit regressions	The level of media attention and the tone of media coverage play an important role in managers' decisions to abandon value-reducing acquisition attempts.
Loughran and McDonald (2013)	S-1 filings	1997-2010	Dictioary-based *L&M*	Linear regressions Logit regressions	Large amounts of uncertain text in an S-1 filing generally lead to more valuation uncertainty and, in turn, higher IPO first-day returns, absolute offer price revisions, and subsequent return volatility (60-day period following the IPO).

References

Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, **59**, 1259-1294.

Baker, M. and J. Wurgler (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, **21**, 129-151.

Black, F. (1986). Noise. *Journal of Finance*, **41**, 529-543.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.

Buehlmaier, M. M. M. (2013). The role of the media in takeovers: theory and evidence. Available at SSRN: http://ssrn.com/abstract=1673164.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, **52**, 57-82.

Carretta, A., V. Farina, E. A. Graziano, and M. Reale (2011). Does investor attention influence stock market activity? The case of spin-off deals. Available at SSRN: http://ssrn.com/abstract=1930861.

Chen, H., P. De, Y. Hu and B. Hwang (2013). Customers as advisors: the role of social media in financial markets. Available at SSRN: http://ssrn.com/abstract=1807265.

Das, S.R. and M.Y. Chen (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*. **53**, 1375–1388.

Davis, A. K., W. Ge, D. A. Matsumoto, and J. L., Zhang (2012). The effect of managerial 'style' on the tone of earnings conference calls. CAAA Annual Conference 2012. Available at SSRN: http://ssrn.com/abstract=1982259.

Davis, A. K., J. M. Piger, and L. M. Sedor (2011). Beyond the numbers: measuring the information content of earnings press release language. AAA 2008 Financial Accounting and Reporting Section (FARS) Paper. Available at SSRN: http://ssrn.com/abstract=875399.

Davis, A. K. and I. Tama-Sweet (2011). Managers' use of language across alternative disclosure outlets: earnings press releases versus MD&A. Available at SSRN: http://ssrn.com/abstract=1866369.

De Franco, G., O. Hope, D. Vyas, and Y. Zhou (2012). Analyst report readability. Rotman School of Management Working Paper No. 1873424. Available at SSRN: http://ssrn.com/abstract=1873424.

DeGennaro, R., and R. Shrieves (1997). Public information releases, private information arrival and volatility in the foreign exchange market. *Journal of Empirical Finance*, **4**, 295–315.

Demers, E., and C. Vega (2011). Linguistic tone in earnings announcements: news or noise? FRB International Finance Discussion Paper 951. Available at SSRN: http://ssrn.com/abstract=1152326.

Doran, J. S., D. R. Peterson and S. M. Price (2010). Earnings conference call content and stock price: the case of REITs. *Journal of Real Estate Finance and Economics*, **45**(2), 402-434.

Durnev, A. and C. Mangen (2011). The real effects of disclosure tone: evidence from restatements. Available at SSRN: http://ssrn.com/abstract=1650003.

Engelberg, J. (2008). Costly information processing: evidence from information announcements. AFA 2009 San Francisco Meetings Paper. Available at SSRN: http://ssrn.com/abstract=1107998.

Engelberg, J., A. V. Reed, and M. C. Ringgenber (2012). How are shorts informed? Short sellers, news, and information processing. *Journal of Financial Economics*, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1535337.

Fama, E. F. and K. French, (1992). The cross-section of expected stock returns. *Journal of Finance*, **47**, 427-65.

Fama, E. F. and K. French, (1993). Common risk factors in the returns on bonds and stocks. *Journal of Financial Economics*, **33**, 3-56.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, **49**, 283–306.

Feldman, R., S. Govindaraj, J. Livnat and B. Segal (2008). The incremental information content of tone change in Management Discussion and Analysis. Available at SSRN: http://ssrn.com/abstract=1126962.

Ferguson, N. J., J. Guo, H. Y. T. Lam, and D. Philip (2012). Media content and stock returns: The predictive power of press. Available at SSRN: http://ssrn.com/abstract=2111352.

Ferris, S. P., G. Q. Hao, and M. Liao (2012). The effect of issuer conservatism on IPO pricing and performance. *Review of Finance*, Forthcoming. Available at SSRN: http://ssrn.com/abstract=2037942.

Garcia, D. (2012). Sentiment during recessions. *Journal of Finance*, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1571101.

Glosten, L., R. Jagannathan, and D. Runkle (1993): On the relationship between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, **48**, 1779–1801.

Hart, R. DICTION 5.0. See http://rhetorica.net/diction.htm.

Henry, E. (2006). Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting*, **3**, 1-19.

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

Henry, E. and A.J., Leone (2009). Measuring qualitative information in capital markets research. Available at SSRN: http://ssrn.com/abstract=1470807.

Huang, X., S. H. Teoh, and Y. Zhang (2011). Tone management. Available at SSRN: http://ssrn.com/abstract=1960376.

Huang, A., A. Y. Zang, and R. Zheng (2013). Large sample evidence on the informativeness of text in analyst reports. Available at SSRN: http://ssrn.com/abstract=1888724.

Jegadeesh, N. and A. D. Wu (2012). Word power: A new approach for content analysis. AFA 2012 Chicago Meetings Paper. Available at SSRN: http://ssrn.com/abstract=1787273.

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

Larker, D. and A. Zakolyukina (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research,* **50** (2), 495-540.

Lehavy, R., F. Li, and K. Merkley (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Accounting Review,* **86**, 1087-1115.

Li, F., (2006). Do stock market investors understand the risk sentiment of corporate annual reports? Available at SSRN: http://ssrn.com/abstract=898181.

Li, F., (2010). The information content of forward-looking statements in corporate filings - A Naive Bayesian machine learning algorithm approach. *Journal of Accounting Research,* **48**, 1049-1102.

Liu, B. and J. J. McConnell (2013). The role of the media in corporate governance: Do the media influence managers' capital allocation decisions?. Available at SSRN: http://ssrn.com/abstract=2004567.

Loughran, T., and B. McDonald (2011a). When is a liability not a liability?. *Journal of Finance,* **66**, 35-65.

Loughran, T. and B. Mcdonald (2011b). Barron's red flags: Do they actually work? J*ournal of Behavioral Finance*, **12** (2), 90-97.

Loughran, T. and B. Mcdonald (2013a). IPO first-day returns, offer price revisions, volatility, and Form S-1 language. *Journal of Financial Economics*, Forthcoming. Available at SSRN: http://ssrn.com/abstract=2128766.

Loughran, T., and B. McDonald (2013b). Measuring Readability in Financial Disclosures. Available at SSRN: http://ssrn.com/abstract=1920411.

Manning, C., and H. Schutze (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press.

McCallum, A. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. *Working paper*, School of Computer Science, Carnegie-MellonUniversity.

Mitchell, M. L. and J. H. Mulherin (1994). The impact of public information on the stock market. *Journal of Finance,* **49**, 923-950.

Nelson, D.(1991): Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, **59**, 347–370.

Ozik, G. and R. Sadka (2012). Media and investment management. Available at SSRN: http://ssrn.com/abstract=1633705.

Price, S. M., J. S. Doran, D. R. Peterson, and B. A. Bliss (2012). Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *Journal of Banking and Finance,* **36**(4), 992-1011.

Rees, L. and B. Twedt (2012). Reading between the lines: an empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy,* **31**, 1-21.

Rogers, J. L., A. Van Buskirk, and S. L. C. Zechman (2011). Disclosure tone and shareholder litigation. *Accounting Review,* **86**(6), 2011.

Sinha, N. R. (2010). Underreaction to news in the US stock market. Available at SSRN: http://ssrn.com/abstract=1572614.

Stone, P. J., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge: MIT Press.

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

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

Endnotes

[1] The studies we review are confined to those examining the role of positive and/or negative sentiment, or 'uncertain', 'deceptive' affects. Some other studies (e.g. Lehavy et al. (2011), De Franco et al. (2012) and Loughran and Mcdonald (2013b)) which investigate the complexity/readablity of texts, are not included in this survey.

[2] http://finance.yahoo.com.

[3] http://www.ragingbull.com.

[4] http://seekingalpha.com.

[5] http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html.

[6] http://rhetorica.net/diction.htm.

[7] The L&M finance dictionary consists of all words that occur in 5% or more of the 10-Ks from 1994 to 2008 (Loughran and McDonald (2011a)).

[8] http://www2.reuters.com/productinfo/newsscopesentiment.

[9] $UE_t = E_t - E_{t-4}$, $SUE_t = (UE_t - \mu_{UE_t})/\sigma_{UE_t}$, where E_t is the firm's earnings in quarter *t*, UE_t is the unexpected earnings. The mean (μ) and standard deviation (σ) of unexpected earnings are calculated on the basis of previous 20 quarters data.

[10] The log number of small, medium and large trades and the log number of traded shares.

[11] The daily average of the bid–ask spread

[12] i.e. Cash flow from operations in the next four quarters scaled by the book value of current liabilities at the end of that quarter.

[13] The GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model, introduced by Bollerslev (1986), is a process of the conditional variance (volatility) of the dependent variable (stock returns, in this case). The conditional variance is expressed as its own lags and lags of the squared error term. Compared with GARCH, the GJR-GARCH model (Glosten et al.(1993)) also models asymmetry in the volatility process. The EGARCH model (Nelson (1991)) models the process of log conditional variance. In the GARCH-X, GJR-GARCH-X and EGARCH-X models, exogenous variable X is also incorporated in the volatility process.

[14] Excess returns in a long period (e.g. 80 days) following the event.

[15] The difference between quarterly EPS and the mean consensus analyst forecast deflated by stock price at the beginning of the quarter.