
# Lie to Me:

## Using Analytics and Content Analysis to Detect Deception

Recent court rulings in both the United States and Germany will make it more difficult for investors to seek recourse if they believe the company they invested in has deceived them (Norris 2014). For example, a judge in Stuttgart, Germany, ruled that companies do not have the right to lie to their shareholders, but said deception is sometimes necessary (Norris 2014) to hide strategic decisions that are not yet public. To help properly vet companies and reduce the possibility that a company is willingly attempting to mislead or defraud, tools are needed that offer insight into the intentions of a company by identifying deceptive communication patterns. Competitive intelligence professionals also need tools to help determine if a competitor is attempting to deceive the market or hide strategic intentions. Content analysis is an instrument that is helpful to investors and competitive intelligence professionals to reduce uncertainty about company intentions.

**David Krauza**

## WHAT IS CONTENT ANALYSIS?

Content analysis, in its most basic form, is a reading of text and other forms of communication to investigate the character of the message. The emphasis of content analysis is not on what is being communicated, but rather, the nature of how the message is being communicated. Bernard Berelson, the late American behavioral scientist, defined content analysis as "a research technique for the objective, systematic and quantitative description of the content of communication" (1952). Klaus Krippendorff, a professor in the Annenberg School of Communication at the University of Pennsylvania, describes content analysis as "a research technique for making replicable and valid inferences from texts to the contexts of their use" (2004). Both Berelson and Krippedorf are describing content analysis as a tool that is used to understand the meaning of a message.


An analyst using content analysis will be interested in word use, description of topics, consistency of word usage, and connection of words (Neuendorf 2002). The ultimate goal of content analysis is to arrive at the underlying theme of the communication. A key advantage of content analysis is that the technique is unobtrusive, meaning the researcher can analyze content and the role the communication plays in the lives of the sender and receiver without the parties involved being aware of the analysis (Krippendorff 2004).

## HOW DO WE DEFINE DECEPTION?

Communication researchers define deception as false communication that tends to benefit the communicator (Mitchell 1986). They say deception is an act that is intended to foster in another person a belief or understanding, which the deceiver considers to be false (Krauss 1981). For our purposes we will define deception as "a successful attempt to plant a false belief in another individual, without the individual being forewarned, that the sender of the message knows to be false for the benefit of the sender." This definition assumes that deception is for the benefit of the message sender. The definition also implies that deception is a malicious act but not necessarily one in which the sender will receive a direct financial benefit.

## BACKGROUND OF THE PROBLEM CONTENT ANALYSIS WILL SOLVE

At the end of 2011, the Federal Bureau of Investigation was actively pursuing 726 corporate fraud cases (FBI 2011). These cases involved losses to investors exceeding $1 billion (FBI 2011). In 2011, the FBI secured $2.4 billion in restitution orders and $16.1 million in fines from corporate criminals (FBI 2011). The number of corporate fraud cases the FBI has pending has increased every year since 2007. Also, since 2006 there has been a general increase in the number of Federal Securities Class Action Law Suits filed against companies alleging fraud (Securities Class Clearinghouse 2012).

In any given year it is estimated that up to 7% of firms commit some act of fraud with less than 2% of these firms being caught (Dyck, Morse and Zingales 2007). These fraudulent activities cost companies nearly 3% of their enterprise value (Dyck, Morse and Zingales 2007). The tools currently used to catch fraudulent corporate acts have limited effectiveness. The institutions and organizations that have been put into place to protect investors do a poor of job identifying fraudulent activity. In an average year, in the United States, the Securities and Exchange Commission identifies roughly 7% of the companies engaged in fraud; corporate auditors only do slightly better, identifying about 10% of the companies (Dyck, Morse and Zingales 2010).

The failure to predict and detect corporate fraud is very costly to investors and the economy as a whole. For example, Enron's bankruptcy in November 2001 caused its shareholders, collectively, to lose $1 billion. However, what cannot be measured from scandals such as Enron is the psychological cost to the economy. If investors, market participants, and other businesses can no longer trust their counter-parties in business transactions, economic activity will decrease, if not completely dry up. Content analysis may help provide indicators of potentially fraudulent or deceptive activity at a company before the effects of the malfeasance manifest on the financial statements of the firm or disrupt competitors.

## WHY WILL COMPANIES ATTEMPT TO DECEIVE?

Pamela Meyer, founder and CEO of Calibrate, a leading deception detection training company based in Washington DC, said, "[Deception] is an attempt to bridge a gap, to connect our wishes and our fantasies about who we wish we were, how we wish we could be, with what we are really like" (2011). When a company attempts to deceive its shareholders or competitors, the company is attempting to tell a story of what they wish they could be and what they wish the market conditions could be like.

Typically, people will behave dishonestly enough to profit from the deception, but honestly enough to delude themselves of their own integrity (Mazar, Amir and Ariely 2008). Nina Mazar of the University of Toronto, showed that if people fail to comply with their internal standards for honesty, they will need to update their view of themselves (2008). This leads people to engage in malleable behaviors that allow for the reinterpretations of their actions in a self-serving manner (Mazar, Amir and Ariely 2008). Other researchers, such as Gordon, Miller, Bond, and DePaulo, have found that deceivers often conclude that their deceptive actions were justified and that they had, in fact, not done anything wrong. The implication of this means that individuals who engage in deceptive practices will attempt to structure and justify their actions in such a way that they do not believe they are engaging in untoward activities.

Social science researchers, Schrand and Zechman, say business executives who are overconfident and exhibit an optimistic bias are more likely to make decisions that could lead to deceptive financial reporting. The researchers found a positive correlation between executive overconfidence and the SEC issuing an Accounting and Auditing Enforcement Release (AAER) and requiring a restatement of financial statements.

According to Paul Tetlock of the Columbia Business School, high levels of media pessimism predict downward pressure on a firm's stock price, and in the case of a small firm, the downward movement of the stock price is sticky, meaning it will not rebound when pessimism disappears (2007). The media pessimism and resulting downward pressure on stock prices could potentially provide the management of a company a reason to deceive the media about the firm's actual results in an attempt to raise the stock price.

## WHAT DO THE EXPERTS SAY, CAN YOU REALLY DO THIS?

Several linguistic researchers (Newman, et al. 2003, Pennebaker 2011) have demonstrated that features of linguistic style, such as pronoun use, emotionally toned words, and prepositions and conjunctions that signal cognitive work, are linked to a number of behavioral and emotional outcomes, including deception. Across five studies performed by Matthew Newman of the University of Texas, deceptive communications were characterized by fewer first person singular pronouns, fewer third person pronouns, more negative emotion words, fewer exclusive words, such as "without" and "but", and more motion verbs (2003). Newman found that deceivers tend to use fewer exclusive words and used third person pronouns at a lower rate than truth-tellers (2003). Newman also found that deceivers tend to tell stories that are less complex, less self-relevant, and more characterized by negativity (2003).

Julia Hirschberg of Columbia University wrote that deceivers use more passives, negations, and indirect speech than truth-tellers (2010). She goes on to write that deceivers provide fewer details, exhibit less cognitive complexity in their speech, and stray from the topic more frequently by mentioning peripheral events or relationships (2010). Hirschberg said deceivers tended to make significantly more negative statements and complaints and repeat words and phrases more often than truth-tellers (2010).

## HAS THIS EVER BEEN USED BEFORE?

For decades, analysts at the Central Intelligence Agency and the Department of Defense have compiled psychological assessments of hostile foreign leaders (Carey 2011). The profiles made use of content analysis techniques. Among the tools used by the intelligence specialist is a software program developed by Margaret Hermann, the director of the Institute of Global Affairs at Syracuse University, that evaluates the relative frequency of certain categories of words (like "I," "me," "mine") in interviews, speeches, and other sources and links the scores to leadership traits (Carey 2011).

The intelligence agencies also leverage a technique developed by David G. Winter, a professor of psychology at the University of Michigan, that judges a leader's motives, in particular their need for power, achievement, and affiliation (Carey 2011). For example, the sentence, "We can certainly wipe them out," reflects a high power orientation; the comment, "After dinner, we sat around chatting and laughing together," rings of affiliation.

## IDENTIFYING COMPANIES ATTEMPTING TO DECEIVE

I undertook a study to test previous research and to examine the open source content published by publicly traded companies to find patterns that indicate patterns of deceptive communication. My study used the classic experiment approach with an experiment and control group. To find companies that engaged in fraudulent or deceptive behavior for my experiment group, I looked at SEC Accounting and Auditing Enforcement Releases (AAER). An AAER is a document published by the SEC at the conclusion of an investigation into a company, an auditor, or an officer of a company for alleged misconduct. I examined AAERs that were published between August 1, 2002 and July 31, 2012. During this time period a total of 1,793 AAERs were released. I scanned each of the AAERs by searching for the term "fraud" and derivatives of the word "fraud" in the text of the release. The SEC uses very direct language in their releases. If the SEC believes fraudulent activity has occurred, the agency will call it fraud. Examples of the language the SEC uses in AAERs include: "committed securities fraud in accounting for certain software agreements" (SEC 2004) and "fraudulent accounting practices designed to inflate its reported revenue" (SEC 2005).

My initial scan of AAERs covered 170 different SIC codes. To make the amount of content manageable for my study, I decided to focus just on the software industry. The AAER also contains very helpful information, besides the description of the offense. The AAER specifically indicates the beginning and ending dates for the alleged fraud. For example, "[F]or the four years ended December 31, 2001 and the first three quarters of 2002, i2 misstated approximately $1 billion of software license revenues" (SEC 2004).

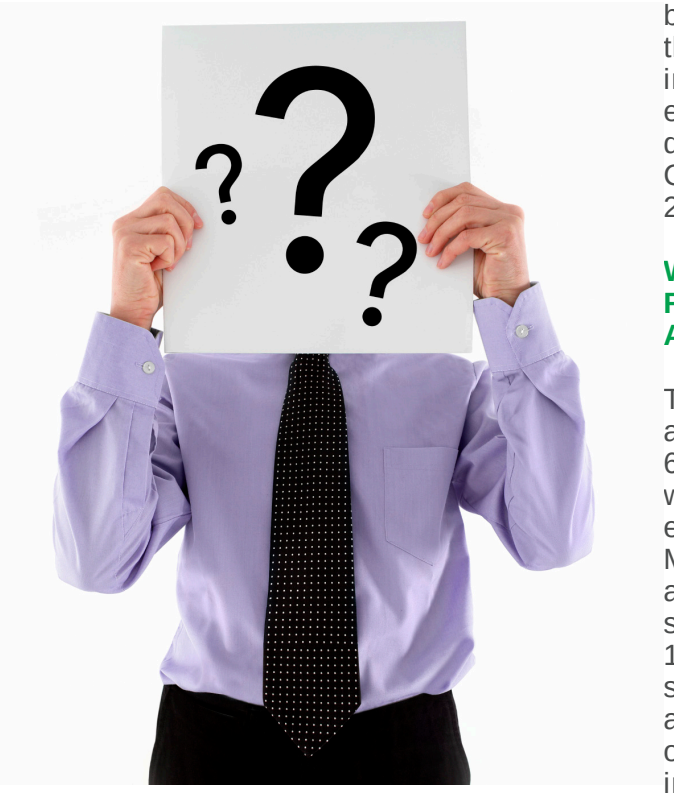
This information was important to help me define the control group of firms for this study. I picked companies for my control group at random from a list of companies with the same SIC code as the companies that were accused of fraud. When a company was selected for entry into the control group, the only criteria the company had to meet was that it had not been accused of fraud contemporaneously with the deceptive companies.

Data for my study was collected indirectly by gathering existing documents that the deceptive and control companies produced. The key advantage to using pre-existing content is that the texts are actual real-world work products. The types of documents

I collected included regulatory filings, such as 10-Ks and 10-Qs, Annual Reports and Letters to Shareholders, and the Transcripts of Earnings Calls and Media Appearances by executives of the companies I was examining.

## IDENTIFYING PATTERNS OF DECEPTION

To identify patterns of deception I leveraged a financial sentiment dictionary that was developed by Notre Dame Professors Tim Loughran and Bill McDonald. The dictionary was developed to determine the level of negative, positive, negation, litigious, and modal words in a piece of content. Modal words express levels of confidence. Strong modal words are confident, they include words such as "always," "must," and "will." Weak modal words are used to hedge, they include words such as "could," "might," and "possibly." The Financial Sentiment Dictionary also controlled for words that have multiple meanings (Loughran and McDonald 2011).



To save me from having to manually classify 4.7 million words in the content being analyzed, I used WordStat from Provalis Research. WordStat is a software package that allows for the extraction of themes and trends from unstructured text (Provalis Research 2012). It allows for the use of sentiment dictionaries that are customized for the type of text being analyzed (Provalis Research 2012). WordStat allowed me to calculate the percentage of words that carry negative, positive, uncertain, litigious, and modal meaning as they were defined in the financial sentiment dictionary. In addition to the word usage, the software also calculated the correlation coefficient in each of the word categories (e.g. negative, positive) to determine if there is a relationship between word category usage and either of the two groups of content.

I also leveraged the linguistic research conducted by James Pennebaker of the University of Texas. Pennebaker's work indicates that the way a person uses pronouns can indicate attempts to deceive. The type of language deceivers will use is also different. People who are attempting to hide their intentions or emotions will use relatively simple language, smaller words, shorter sentences, and fewer cognitive words (2011).

Again, to save me from having to manually identify each pronoun in the content and assign a classification to each word, I used the Linguistic Inquiry and Word Count (LIWC) software tool. LIWC was designed by James Pennebaker, Roger J. Booth, and Martha E. Francis (Pennebaker Conglomerates, Inc. 2012). The tool analyzes written text on a word-by-word basis, calculating the percentage of words in the text that match each of up to 82 language dimensions (Pennebaker Conglomerates, Inc. 2012).

## WHAT WERE THE RESULTS OF THE ANALYSIS?

The source data I analyzed comprised 602 source documents which represents content extracted from the Management Discussion and Analysis (MD&A) section of SEC Forms 10-K and 10-Q, letters to shareholders and, when available, transcripts of earnings calls with investment analysts. A total of 4.7 million words were contained in all of the cases analyzed. The documents were created by the companies from June 1993 to June 2011 and, as I stated before, were sorted into an experiment group, containing companies engaging in deception, and a control group.

The output of the analysis using the Loughran and McDonald Financial Sentiment Dictionary showed that negative words comprised the greatest percentage of the key words from the dictionary found in the test cases. Negative words, uncertainty words, and modal words weak accounted for a total of 72.8% of keywords found in the test cases. Negative words accounted for 29.3%, uncertainty words accounting for 28.9%, and modal words weak accounted for 14.6% of the keywords found in the test cases. Figure 1 shows the total percentage breakdown of keywords found in the test cases. accounted for 14.6%of the keywords found in the test cases. Figure 1 shows the total percentage breakdown of keywords found in the test cases.

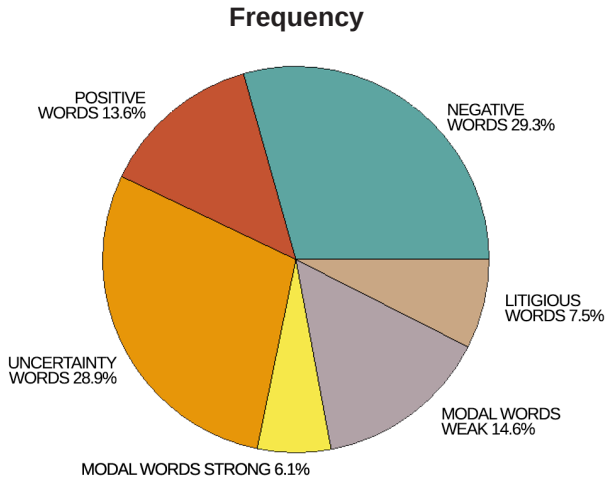



**Figure 1:** Financial Sentiment Keywords Breakdown

However, when digging deeper, the data indicated that companies in the deception group used keywords from the negative, uncertainty, and modal weak word categories as a greater percent of their total words than companies from the control group. On the surface, these results tend to confirm previous research, which states that deceivers would use more non-extreme, negative, and passive words. To determine if the differences in word category usage are meaningful and predictive between the companies in the fraud and control groups, the differences were tested for statistical significance. Figure 2 shows the results of the test.

| Keyword Category | Fraud | Control | P (1-Tail) |
|---|---|---|---|
| **Negative Words** | 2.020% | 1.913% | 0.016 |
| **Uncertainty Words** | 1.978% | 1.891% | 0.006 |
| **Modal Words Weak** | 1.054% | 0.915% | 0.066 |
| **Positive Words** | 0.914% | 0.902% | 0.020 |
| **Litigious Words** | 0.534% | 0.476% | 0.010 |
| **Modal Words Strong** | 0.406% | 0.409% | 0.001 |

**Figure 2:** Probability of Obtaining a Test Statistic

These results indicate that companies in the deception group are statistically more likely to use negative, uncertainty, and litigious words in the content they produce. It also indicates that companies in the control group are statistically more likely to use modal strong words than companies in the deception group. Interestingly, the results also indicate that the deception companies are more likely than the control companies to use positive words in the content generated. It is likely that significant higher use of positive words by the deception companies is related focused on the possibilities that the future may hold rather than current problems.

Overall, the word count data results support Pennebaker (2011), noting that deceivers will use less words and shorter sentences. The word count data showed that the companies in the deception group used an average of 6,477 words in their content, while the control companies used an average of 8,394 words in their content. The control companies used almost 130% more words in the content pieces than the companies in the deception group. However, to adjust for the possibility that one of the exceptionally large or small case may skew the results, the median values of both groups were also compared. This comparison showed that the deceptive companies used a median of 5,791 words while the control companies used a median of 6,790 words. Even using the median values the control companies used almost 120% more words than the deceptive companies.

Figure 3 shows the distribution of total word usage across all cases. From the diagram it is apparent that the companies in the deception group are tightly clustered toward the lower word count totals. The deceptive group also lacks cases in which more than 22,000 words were used. The control group shows a tight cluster at a much higher word count than the deceptive group and has several cases that exceed 25,000 words.

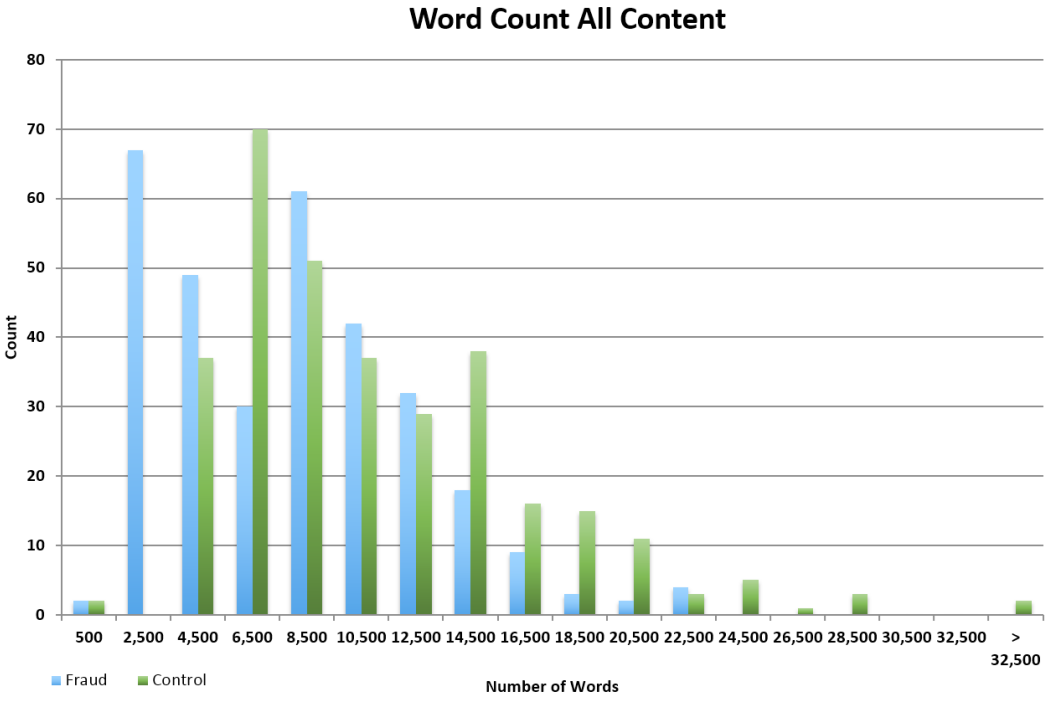



**Figure 3:** Total Word Count Distribution

The sentences in the content of the companies in the deceptive group were generally shorter than the sentences used in the content of the control group. However, the difference is small. Companies in the deceptive group used an average of 26.68 words per sentence in their content versus the control companies, which used an average of 27.63 words per sentence. Figure 4 shows a comparison of the distribution of words per sentence for deceptive and control group. From this exhibit it can be seen that while the mean for both groups is close, the control group has a wider dispersion of sentence length than the deceptive group companies.

## HOW CAN I USE THIS INFORMATION?

The findings contained in this study are easily reproducible by competitive intelligence professionals who are interested in examining companies to find indicators that the company may be engaging in fraud or deception. Practitioners are able to examine the content of companies over any time period that is necessary. The data required to undertake similar studies is, for the most part, readily available on the Internet.

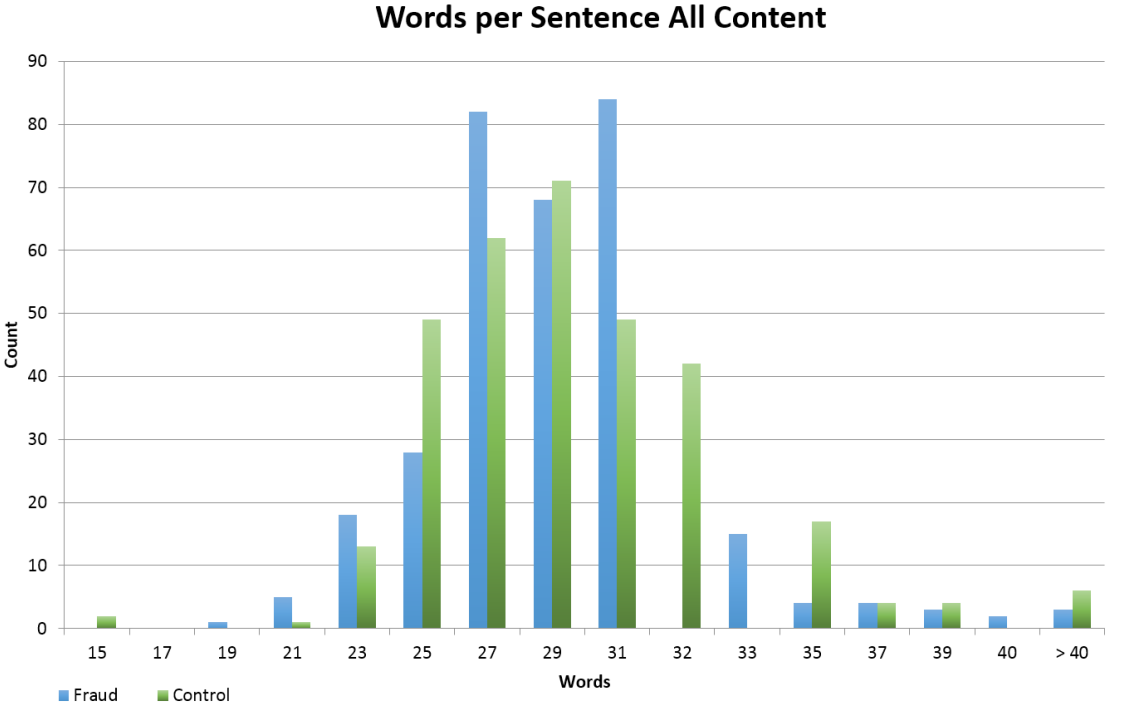



**Figure 4:** Words per Sentence Distribution

## INVESTMENT MANAGEMENT

Members of the investment and asset management community have historically employed analytical methods centered on the analysis of financial statements or the use of statistical and mathematical models to identify potential investments. It was observed during the data collection for this study that there could be a significant lag in time between the manifestation of fraud on the financial statements or the recognition by a regulatory body. Content analysis provides a tool to help investment practitioners to reduce uncertainty around a potentially fraudulent company and identify them sooner to avoid them as investment targets.

## REGULATORY ENFORCEMENT AGENCIES

Regulatory enforcement agencies will be able to find a use for the results found in this study. Regulatory agencies, such as the SEC, account for the identification of only 7% (Dyck, Morse and Zingales 2010) of corporate fraud. Employing the method described in this study will allow agencies to proactively identify more potential cases of fraud and thereby reduce the impact fraudulent companies may have on the overall economy.

## FINAL THOUGHTS

My study demonstrated that indicators of deception exist in the open source content published by publicly traded companies. It indicates that there is a relationship between the words used by companies that are engaged in deception and companies that are not. The methodology employed in this study allows practitioners to assess the risk of being a victim of a fraudulent or deceptive business transaction.

However, the results found in this study are not the only means to determine if a firm is engaged in deceptive activity. This study provides only one potential technique to reduce uncertainty and surprise when looking for indicators of fraud or deception. The results of this study should be considered in light of what is occurring in the macroeconomic environment. It is likely that the results found in this study will be strengthened when the methodology employed in this case is used in conjunction with other analytic methodologies.

## REFERENCES


Berelson, Bernard. 1952. Content Analysis in Communications Research. New York, NY: Free Press. Bond, Charles F, and Bella M DePaulo. 2006. "Accuracy of Deception Judgements." Personality and Social Psychology Review 10 (3): 214-234.

Carey, Benedict. 2011. "Teasing Out Policy Insights From a Character Profile." New York Times, March 28. http://www.nytimes.com/2011/03/29/science/29psych.html?pagewanted=all&_r=0.

Dyck, Alexander, Adair Morse, and Luigi Zingales. 2007. "How Pervasive is Corporate Fraud?" 2nd Annual Conference on Empirical Legal Studies.

Dyck, Alexander, Adair Morse, and Luigi Zingales. 2010. "Who Blows the Whistle on Corporate Fraud?" Journal of Finance 65 (6): 2213-2253.

FBI. 2011. Financial Crimes Report to the Public. September 30. Accessed December 9, 2012. http://www.fbi.gov/stats-services/publications/financial-crimes-report-2010-2011.

Gordon, Anne K, and Arthur G Miller. 2000. "Perspective Differences in the Construal of Lies: Is Deception in the Eye of the Beholder?" Personality and Social Psychology Bulletin 26 (1): 46-55.

Hirschberg, Julia. 2010. "Deceptive Speech: Clues from Spoken Language." In Speech Technology: Theory and Applications, edited by Fang Chen and Kristiina Jokinen, 79-88. New York, NY: Springer Science+Media.

Krauss, R. M. 1981. "Impression Formation, Impression Management, and Nonverbal Behaviors." Edited by E.T. Higgins, C.P. Herman and M.P. Zanna. Social Cognition: Vol 1. The Ontario Symposium. Hillsdale, NJ: Erlbaum. 323-341.

Krippendorff, Klaus. 2004. Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA: Sage Publications.

Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." The Journal of Finance 66 (1): 35-65.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." The Journal of Marketing Research 45: 633-644.

Meyer, Pamela. 2011. "How to Spot a Liar." TED. July.

Accessed September 25, 2012. http://www.ted.com/talks/pamela_meyer_how_to_spot_a_liar.html.

Mitchell, Robert W. 1986. "A Framework for Discussion Deception." In Deception, Perspecitives on Human and Nonhuman Deceit, edited by Robert W Mitchell and Nicholas S Thompson. Albany, NY: State University of New York, Albany.

Neuendorf, Kimberly A. 2002. The Content Analysis Guidebook. Thousand Oaks, CA: Sage Publications.

Newman, Matthew L, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. "Lying Words: Predicting Deception From Linguistic Styles." Personality and Social Pyschology Bulletin 29 (5): 665-675.

Norris, Floyd. 2014. "Corporate Lies Are Increasingly Immune to Investor Complaints." New York Times. March 20. Accessed March 2014. http://www.nytimes.com/2014/03/21/business/corporate-lies-are-increasingly-immune-to-investor-complaints.html?_r=0.Pennebaker Conglomerates, Inc. 2012. LIWC: Linguistic Inquiry and Word Count. Accessed April 15, 2012. http://www.liwc.net.

Pennebaker, James W. 2011. The Secret Life of Pronouns: What Our Words Say About Us. New York, NY: Bloomsbury Press.

Provalis Research. 2012. Content Analysis Software. Accessed November 26, 2012. http://provalisresearch.com/products/content-analysis-software/.—. 2012. WordStat Dictionary. Accessed November 26, 2012. http://provalisresearch.com/products/content-analysis-software/.

- 2012. WordStat Dictionary. Accessed November 23, 2012. http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/.

Schrand, Catherine, and Sarah Zechman. 2012. "Executive Overconfidence and the Slippery Slope to Financial Misreporting." Journal of Accounting and Economics 53 (1/2): 311-329.

SEC. 2004. "Accounting and Auditing Enforcement Release 2034." Securities and Exchange Commission. July 9. Accessed October 14, 2012. http://sec.gov/litigation/litreleases/lr18741.htm.

- 2005. "Accounting and Auditing Enforcement Release 2249." Securities and Exchange Commission. July 9. Accessed October 15, 2012. http://www.sec.gov/litigation/litreleases/lr19260.htm.

Securities Class Clearinghouse. 2012. Securities Class Action Law Suit Clearing House. December 3. Accessed December 9, 2012. http://securities.stanford.edu/.

Tetlock, Paul. 2007. "Giving Content to Investor Sentiment:
The Role of Media in the Stock Market." Journal of Finance 62 (3): 1139-1168.

---

*David is a competitive intelligence professional and consultant based in Philadelphia. He specializes in helping organization develop competitive insights and translating them into actionable strategies. David can be reached at david.krauza@outlook.com.*