On Using Text Analytics for Event Studies*

Max R. Kimbrough, Esquire
Del Collo & Mazzanti LLP
Paoli, PA 19301
mrk@delmazz.com

Steven O. Kimbrough
University of Pennsylvania
Philadelphia, PA 19104
kimbrough@wharton.upenn.edu

Priscilla Murphy
Temple University
Philadelphia, PA
murphyp@temple.edu

ABSTRACT

Event studies seek convincing evidence connecting behavior with a known or conjectured event. Originating in finance, event studies have spread to and become established in many other fields, including law. Multiple regression modeling, the standard methodology for doing event studies, may not be ideally suited for event studies based on data derived from bodies of text, since the required distributional assumptions may be problematic. This paper reports on an exploratory study that uses text analytic methods to discern events. The study draws upon and extends a previously published study of the speeches made by the CEO of a tobacco company in the years surrounding the tobacco settlement in 1998. Examining SEC filings from three major tobacco companies, the study reported in this paper finds cause for optimism in the use of text analytics to discern and investigate events of interest in law.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Indexing methods

Keywords

text, text analytics, event study, law, evidence

1. BACKGROUND: EVENT STUDIES

A prototypical event study comprises:

1. A known event, occurring at a known time and presumed to carry information

---

*(Produces the permission block, and copyright information). For use with SIG-ALTERNATE.CLS. Supported by ACM.

2. A response variable, hypothesized to be influenced by the event

3. Data and analysis, usually in the form of fitting an econometric (statistical regression) model, aimed at determining whether or not the hypothesized response is real (statistically significant) and if so, what its magnitude is.

Such econometric event studies originated and found success in the field of finance [10]. There, the response variable of interest has been some measure of "abnormal" (unusual, unexpected) returns, usually as a function of stock prices. Events (event types) of interest have included merger announcements, acquisition announcements, stock splits, hostile takeovers, and firings of CEOs. For example, it is of interest in finance to know whether being sued by the government is detrimental to a firm's value as measured by its stock prices (it is); whether being sued by another firm for contract violations is detrimental (it is, but not so much); and so on.

The basic methodology for doing this kind of event study is by now long established and settled [5, 9]. Its considerable successes have led to widespread application within finance and to adoption and use in many other fields, including accounting, organizational behavior, business strategy, marketing (e.g., [1]), and law in all its facets, including civil, criminal, administrative, and regulatory, which is of course our focus in this paper.

Event studies in the law have a long history, are actively undertaken today, and the methodology is commonly presented in law school curricula (see [2, 3, 12]). The following passage, from part II of a recent review of event studies in law, illustrates.

> Event studies have had a major impact on corporate law. The explanation for this influence is straightforward. The objective of U.S. corporate law is furthering the interest of the owners of the firm, and the event study methodology, measuring the unexpected change in stock price due to new information about firm value, such as adoption of a new corporate law or a firm decision, provides a metric for identifying whether a specific corporate policy or action has the legal regime's desired beneficial impact on firm owners. [3]

It is to be expected that a well-defined, practicable method (for event studies in their traditional form) that is successful

in one field (finance) will be applied in other fields, such as law. It is also to be expected that abstractions and generalizations of the method will be developed. So it is not surprising that scholars have used event study methods to reverse the chain of inference. Instead of using an observed event (type) to predict abnormal returns, we might use abnormal returns to predict an event type. Or in the following case, we can use the lack of abnormal returns to support an hypothesis on information leakage.

> Shares trading in the Bolsa Mexicana de Valores do not seem to react to company news. Using a sample of Mexican corporate news announcements from the period July 1994 through June 1997, this paper finds that there is nothing unusual about returns, volatility of returns, volume of trade or bid–ask spreads in the event window. We provide evidence that suggests that unrestricted insider trading causes prices to fully incorporate the information before its public release. The paper thus points toward a methodology for ranking emerging stock markets in terms of their market integrity, an approach that can be used with the limited data available in such markets. [4]

In this and similar studies (e.g., [11]), the (observed event, return response) pair, with the event modeled as a predictor of the response in the form of equity returns, is generalized. We now have the triple (observed event, hypothesized event, return behavior). Together, the observed event and the return behavior (both after and *before* the observed event) are used as a basis for an abductive inference to the hypothesized event. Event studies, then, may be run "forward" (to predict) or "backward" (to explain).

Further generalizations are possible and of interest. Our focus in this paper is on generalizing event studies from dealing with only the returns behavior of a firm (variables having to do with its stock price) to dealing with other behavioral variables. In particular, we are interested in examining linguistic behavior and its connection with events of interest. Given an event (observed or hypothesized), here are some characteristic questions of interest:

1. Was there a change in linguistic behavior? (forwards: as a result of the observed event; backwards: that would explain an unobserved event)

2. If there were changes in linguistic behavior, when did they occur?

3. If there were changes in linguistic behavior, what characterizes them? Are there discernible patterns, such as a trend?

4. If there were changes in linguistic behavior, what was it that changed? Was it subject matter? Valence? Frequency of topics mentioned?

In the next section we describe a study that examines the effect of regulatory-legal events on the linguistic behavior, and strategic behavior, of a major firm.

2. THE OLIVEIRA–MURPHY STUDY

2.1 The Tobacco Settlement

In November 1998, the four largest U.S. tobacco companies and the attorneys general of 46 states agreed to what is known as the Tobacco Master Settlement Agreement (MSA). The other four states had already settled with the tobacco companies. This was a culmination of legal action initiated by the Attorney General of Mississippi in 1994 and later joined by 40 states. For years prior to this, individuals had sued tobacco companies without success. The premise of the states' suits was different. The states sought to recover damages from the tobacco companies for costs incurred by them in caring for patients under Medicare.

The agreement constituted a momentous event for the tobacco industry. The signatory firms agreed broadly to the following actions [16]:

1. restrict their advertising, sponsorship, lobbying, and litigation activities, particularly as those activities were seen as targeting youth;

2. generally to make available to the public documents the tobacco companies had disclosed during the discovery phase of their litigation with the settling states;

3. create and fund an institution for public education (American Legacy Foundation) dedicated to reducing youth smoking and preventing diseases associated with smoking;

4. make $368.5 billion in payments to the states during the subsequent 25 years in partial compensation for tobacco-based medical costs.

Big events indeed. Did they affect the linguistic behavior of the tobacco companies?

2.2 The Oliveira & Murphy (O&M) Study

At the time of the settlement, the four largest tobacco companies were PhilipMorris USA (later Altria), R. J. Reynolds Tobacco Company, Brown & Williamson Tobacco Corporation, and Lorillard Tobacco Company, controlling about 97% of the U.S. market.

The largest of these companies was Philip Morris, which had a single CEO from 1994 to 2002, Geoffrey Bible. Quite unusually, and as a result of the settlement (see item 2 in the previous section), all of his speeches during this period became publicly available. Oliveira and Murphy [13] located 67 speeches given by Bible over the period of his tenure. These speeches were given to one of three audiences: employees (23), investment analysts (40), and Congressional committees (4). Except for the four speeches to Congressional committees, none were publicly available documents, or rather would not have been publicly available absent the governing disclosure process. Oliveira and Murphy analyzed the speeches by year, by audience, and as an entire group. They examined how the CEO's speeches portrayed Philip Morris to its internal and external publics, and whether the speeches' themes and word choice altered during this period to reflect the company's social and legal context. Their two main research questions were:

RQ1: How did the CEO's speeches portray Philip Morris to its internal and external publics?

RQ2: Did the speeches' themes and word choice alter over this period to reflect the company's social and legal context?

2.3 CRA: Centering Resonance Analysis

Oliveira and Murphy addressed their research questions by deriving text analytics data from the 67 speeches given by Philip Morris's CEO. They relied primarily on the centering resonance analysis (CRA) method, as implemented in the Crawdad text analysis software product, available from Crawdad Technologies, LLC.[1] Descriptions of CRA are available in the open literature [6, 7, 8]. In a nutshell, CRA is a method of computer-assisted, network-based text analysis that assumes texts achieve coherence through conversation "centers" consisting of nouns or noun phrases. These noun phrases link to form a network of semantic meaning that offers a fundamental representation of the underlying text in abbreviated form; it also shows which words and concepts are most influential in structuring its meaning. One product of CRA is a set of "influence scores" for nouns and noun phrases, which can then be further analyzed with such methods as cluster analysis to group texts according to thematic similarities. An additional feature of CRA is its ability to track changes or compare meanings that evolve over time. As a result, centering resonance analysis enables longitudinal study of the evolution of tobacco industry issues management strategies from 1994–2002.

2.4 Summary of O&M's Findings

Oliveira and Murphy report a number of findings [13], but for our purposes the most important of these have to do with their temporal analysis of the speeches. They grouped the speeches by year, combined the speeches for a given year into a single document, performed CRA indexing on the resulting 8 documents (1994-2001), and used Crawdad's clustering procedure. Their analysis found three major clusters: (a) speeches from 1994, 1995, and 1996, focusing on the company's growth and international expansion; (b) speeches from 1997 and 1998, portraying the importance of litigation issues and legal restrictions during these years; and (c) speeches from 1999, 2000, and 2001, presenting the tobacco industry's adaptation to a social and business environment transformed by the Master Settlement Agreement [13, pages 370–1].

3. FURTHER RESEARCH QUESTIONS

Oliveira and Murphy conducted an exploratory event study, using text analytics (based on CRA) to produce the data upon which the findings were based. The event study was descriptive in nature, rather than predictive or explanatory, as is appropriate given their research questions. Its descriptive success, however, prompts many additional *general* research questions, having to do with text analytics for event studies, and reaching well beyond the scope of their original study. Examples include:

Q1: Is it generally possible to detect changes in patterns of linguistic usage from publicly available documents?

The O&M study relied on CEO speeches, which are not normally available to the general public. It would be useful to have a body of case studies that look into

the productivity in this regard of documents that are more generally accessible. Of course, whether a change can be detected will depend upon the size of the effect, the extent of the document data base, and the text analytic instruments used for detection.

Q2: Which text analytic methods are effective, and under what conditions are they effective, for discerning changes in patterns of linguistic usage?

CRA is a credible and established approach, but it hardly exhausts the armamentarium of text analytics. We will introduce below a distinct, complementary technique. A full addressing of this question is hardly possible in a single paper; it must be the subject of a larger stream of research.

Q3: Can changes in linguistic usage patterns, detected by text analytics methods, be acceptable as evidence in a legal proceeding, and if so, under what conditions? In particular, among other things we would like to investigate with regard to text analytics are:

1. Whether and if so how text analytics can support the prediction of an otherwise unobserved event. Standard event studies have been used, as we saw above, as evidence for insider trading. What kinds of events might be predicted (or postulated) effectively with text analytics methods?

2. Whether and if so how linguistic usage patterns can be predicted by usage patterns followed (or proceeded) by an event of a given type?

3. What the proper principles are for interpreting discerned changes in patterns of linguistic usage and assigning causes to them. The O&M study interpreted the period (c) speeches as indicating a corporate strategy:

> Strikingly, the speeches [in this period, 1999–2001] address good citizenship as directed toward the community, employees, shareholders—everybody except its own customers. ... the decision not to talk about health issues in Geoffrey Bible's speeches indicates a strategic adaptation to the changed opinion climate. It [Philip Morris] had quietly given up contesting health arguments—clearly a losing strategy—and was acutely aware of the importance of legitimacy, seeking to authenticate its right to do business by establishing other frames, other terms, for judging its contributions to society. [13, pages 376–7]

As a start on addressing these and related questions, we undertook an exploratory study extending the O&M study. Our discussion of that begins in the next section.

4. EXPLORING SEC FILINGS

We created a *categorized document base* (CDB), which we used for our study. Our CDB consists of SEC[2] filings (10-Ks, annual reports; and 10-Qs, quarterly reports) from

---

[1]http://www.crawdadtech.com/

[2]http://www.sec.gov

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	1078	1474	1921	1220	2398	1937	2693	2887	4230
Altria10Qs:	4913	3479	6023	13992	24405	9222	13752	11377	13717
Lorillard10Ks:	10202	6231	4734	6479	6059	6653	6331	32587	20257
Lorillard10Qs:	3261	4143	5911	8583	22006	8061	8158	10700	13106
RJReynolds10Ks:	1782	2489	2740	3093	3359	3709	4905	3999	5203
RJReynolds10Qs:	4575	24923	7598	6728	19497	10733	11170	5681	6988

Table 1: Raw Scores. Query: C:the

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	182.9	187.0	176.2	171.8	154.2	151.3	163.3	157.2	170.7
Altria10Qs:	129.7	106.7	129.9	143.9	131.7	126.0	138.7	131.3	135.9
Lorillard10Ks:	72.99	45.80	111.9	120.0	122.6	124.9	123.6	135.5	132.5
Lorillard10Qs:	162.1	167.3	177.6	174.4	158.8	165.3	177.7	176.3	180.9
RJReynolds10Ks:	170.2	167.1	172.0	164.4	153.4	153.8	148.3	161.5	170.5
RJReynolds10Qs:	149.8	171.2	173.8	177.8	161.1	180.6	144.5	161.1	144.0

Table 2: Normalized Scores. Query: C:the

the reporting years 1994–2002, for the three major tobacco companies operating in the United States and publicly held: Altria (Philip Morris), Lorillard, and RJ Reynolds.[3] Each document in the CDB is classified in three ways, that is along three dimensions:

1. Reporting year (1994–2002)[4]

2. Company (Altria (Philip Morris), Lorillard, or RJ Reynolds)[5]

3. Report type (10-K or 10-Q)

The effect is to create what we call a *text cube* by way of analogy to standard data cubes (or OLAP cubes). We then developed a series of concept representations—sometimes simply a single word, normally a more complex representation—which we matched to the documents in the CDB to get numerical matching scores. (See below for details.) We then report these scores in data cube format. The data cube examples we show here have been reduced to 2-dimensional tables by combining the company and report type dimensions, but without loss of information, as we shall see.

Our primary objective is to explore—to compare and contrast—our results on the SEC filings using our text analytics techniques with the O&M findings on the CEO speeches using CRA. To this end, we present and discuss a series of tabular reports.

We begin with a validation report, which also serves to illustrate our statistical method. Tables 1 and 2 show associations between the concept *C:the* and the articulated categories of our CDB. (We use the notation *C:⟨word-or-phrase⟩* to indicate a concept *C* named by *word-or-phrase*.) Each

cell in the tables presents a numerical score of the measured association between the concept C:the and a triple categorization consisting of a year, a company, and a report type. For example the 1078 in the upper left-hand corner of Table 1 is the raw (unnormalized) association score for C:the and (Altria, 10K, 1994). Looking rightwards in the row, we find a broadly increasing trend over time. The 182.9 in the upper left-hand corner of Table 2 is the normalized (by document length) association score for C:the and (Altria, 10K, 1994). Looking rightwards in the row, we do not see much of a trend. What is the C:the concept? What does it represent? It counts occurrences of the whole words *the, and, a,* and *an.* We would expect that the scores for this (test, validation) concept should *not* change over time, when normalized by the length of the documents. Does it?

Here and in the sequel we consider three periods: period 1 is 1994–1996, period 2 is 1997–1999, and period 3 is 2000–2002. To compare periods 1 and 3 we subtract the sub-table for period 1 from the sub-table for period 3 and record the signs (+ or −) of the cell differences. Comparing periods 1 and 2, and 1 and 3, for Table 2 yields:

Period 1 v. Period 2.

-	-	-
+	+	-
+	+	+
+	-	-
-	-	-
+	-	+

Period 1 v. Period 3.

-	-	-
+	+	+
+	+	+
+	+	+
-	-	-
-	-	-

Under our null hypothesis of no difference between periods 1 and 2, and 1 and 3, we expect the number of +s in the difference tables (above) to equal the number of −s. This happens to be met exactly in the 1–3 comparison. For the 1–2 comparison there are 8 +s and 10 −s. What, then, is the probability of getting 8 or fewer +s if the null hypothesis is correct? If this probability is low enough, standardly below 0.05, then we have grounds to reject the null hypothesis. The relevant probability distribution function is the binomial.

---

[3]We removed the Exhibits sections from the 10-Ks because they were very long and unwieldy, consisting of repetitious boilerplate, unlikely to be informative for our purposes.

[4]The O&M study used the years 1994–2001 because the CEO Bible left Altria after 2001 and his speeches ended. We added 2002 because we thought it would be relevant and it allows us to bracket the 1997–1999 period, with three years on either side. This is helpful for statistical purposes.

[5]The fourth member of Big Tobacco was the Brown & Williamson Tobacco Corporation, but it was privately held and does not have SEC filings.

We want the cumulative distribution value for 8 or fewer +s. For MATLAB fans this is

```
>> binocdf(8,18, 0.5)
ans =  0.4073
```

Our (nonparametric) sign test gives us no reason to reject the null hypothesis. From Table 1 it is evident that the SEC filings are getting longer over time, but as we would hope, the rate of usage of *the, and, a,* and *an* is stable.

Tables 3 and 4 show associations between the concept *C:tobacco products* and the articulated categories of our CDB. Points arising:

1. Our raw concept representation is quite simple. It amounts to just the list of these terms: *tobacco, cigarette, cigarettes, cigar, cigars.* The raw score is the total number of matches to any of these terms in the documents for a category. The normalized score is the raw score divided by the length of the document (here measured as the number of characters in the document file).

2. The issue of concept representation design is an important one, mostly beyond the scope of this paper. The point being made here is that even fairly simple collections of terms (expanded by WordNet [14], which we do) will often yield useful and valid results.

3. Which scoring is correct, raw or normalized? The SEC documents filed vary in size considerably, even holding the type and the firm constant. Consequently, the default should be to work with the normalized scores. We note, however, that in exploratory mode it may be very helpful to examine the raw scores as well. It is entirely possible, and possibly of much interest, that a firm is saying more on a given topic, but saying more overall, so that the given topic diminishes as measured by the normalized scores.

4. Can we do statistical tests? We can, but we are not emphasizing that here. For present purposes, we will limit the discussion to basic (but entirely valid) nonparametric sign tests, as discussed for the C:the concept. Again, the O&M study leads us to distinguish three different periods in the 1994–2002 interval: period 1: 1994–6, period 2: 1997–9, and period 3: 2000–2. This permits a nice nonparametric test. If we subtract the cell values in the 2000–2 period from their analogs in the 1994–6 period and record only the signs (+ or −), we would expect under the null hypothesis that the number of + (or −s) should be binomially distributed with $p = \frac{1}{2}$. For the case at hand, Table 4, making this comparison gives yields 5 −s and 13 +s. The probability of getting 5 or fewer minuses under the null hypothesis is 0.0481. So, at the 5% significance level we can say that there is more talk related to tobacco products (as represented) in the third period than in the first period. Comparing periods 1 and 2, 2 is in excess of 1 in 15 of 18 cases. The probability of this (or worse) occurring if the null hypothesis is true is 0.0038. The 2-3 comparison does not produce a significant difference. We note that there may be systematic differences between the 10-Ks and the 10-Qs, but since we are comparing only 10-Ks with 10-Ks and 10-Qs with 10-Qs, we think the pooling is justified.

5. From the above, it would appear that factors associated with the signing of the Tobacco Settlement may have been forcing these tobacco companies to increase their discussions of their specifically tobacco-related products in these regulatory filings.

Turning now to social responsibility, the O&M study suggests, as we have seen, that Philip Morris took a rhetorical turn in that direction in the third period (2000–2). Is this finding born out with SEC filings? Table 5 indicates this is indeed the case. Notice that *every* cell but one in the third period has a value exceeding its analog in the first period. Under the null hypothesis the chance of this happening is 7.2479×10^{-5}, less than once in 13,000 trials.

The O&M study reports that

> The basic product word *tobacco* featured far more prominently in speeches to the external publics than to the internal publics. Other key words in speeches directed at external groups were *business, company, new, product,* and *brand,* indicating that the Philip Morris CEO constantly addressed issues related to the company's profitability, the brands' values, and the increase in market share when talking to Philip Morris investment analysts, who constitute the main audience for the speeches. As the decade wore on and legislative pressure became more intense, the CEO speeches addressed social responsibility, a primary concern of the other external audience, legislators.

Referring now to Table 6, we represented the profitability concept with the terms: *profit*, business*, compan*, new, product*, brand*, value*,* and *market share,* where the asterisk matches to a variety of endings, such as plurals. O&M are reporting on speeches to external audiences, compared to internal speeches. Since all of our documents are SEC filings, we cannot support the distinction; all of these documents are directed externally. As such, we might expect little or no trend; profitability is a perennial issue. Indeed, we find no significant difference between the first and third periods. There is a very slight hint of a difference between periods 1 and 2, with 6 −s the probability under the null hypothesis of have 6 or fewer −s is 0.1189 (neglecting Bonferroni considerations).

O&M found that Philip Morris's CEO discussed health (not safety, but safety figures very little in the SEC documents) less and less as time passed, tacitly conceding the argument to tobacco's critics. We operationalized our health and safety concept in several different ways, with very similar results. The terms behind Table 7 are: *health, safety, wellness, living, well-being, wellbeing.* With 3 −s in the 3 - 1 comparison, there is a strong statistically significant uptick in health-related language between periods 1 and 3 ($p = 0.0038$). We suspect that the language is more defensive in nature, with the tobacco companies increasingly having to inform their investors of financial consequences of tobacco's adverse health effects. Naturally, a CEO facing an external audience would want to accentuate the positive, and just not discuss it.

Finally and inevitably, legal issues. "During 1996, litigation issues became serious" write O&M. In consequence, "*litigation, issue, FDA,* and *leadership* become central in the

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	118	174	218	170	301	262	320	337	409
Altria10Qs:	223	233	485	895	2030	929	963	998	1033
Lorillard10Ks:	149	159	203	422	566	500	366	1566	836
Lorillard10Qs:	103	174	269	736	1556	571	401	458	430
RJReynolds10Ks:	232	314	287	389	435	469	1042	844	981
RJReynolds10Qs:	269	392	319	357	810	962	1634	1120	1348

Table 3: Raw Scores. Query: C:tobacco products

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	20.02	22.07	20.00	23.94	19.35	20.47	19.41	18.35	16.51
Altria10Qs:	5.890	7.149	10.46	9.206	10.95	12.69	9.719	11.51	10.24
Lorillard10Ks:	1.066	1.168	4.800	7.822	11.45	9.388	7.148	6.516	5.470
Lorillard10Qs:	5.121	7.030	8.083	14.95	11.22	11.71	8.738	7.550	5.935
RJReynolds10Ks:	22.16	21.08	18.02	20.68	19.87	19.45	31.51	34.08	32.15
RJReynolds10Qs:	8.809	2.693	7.298	9.436	6.696	16.19	21.15	31.76	27.77

Table 4: Normalized Scores. Query: C:tobacco products

internal [speeches of Bible], mirroring the rise of tobacco litigation issues in this year." Our operationalization of the legal concept is: *legal, litigation, difficult, regulation, propose, resolution, regulatory, issue, restriction, Congress,* and *challenge.* Table 8 shows the results. With 5 −s in the 2 - 1 comparison, there is a significant increase from period 1 to 2 ($p = 0.0481$), but not from 1 to 3 ($p = 0.1189$).

As a final step, and independently from the above analysis, we grouped the financial documents by company and type (10-K, 10-Q) and ran a clustering analysis with Crawdad on each of the six resulting groups, following procedures used by O&M. The result in each case was a good-quality three- or four-cluster solution that closely, though not completely, reproduced the O&M grouping. For example, clustering the Altria 10-Qs, Crawdad finds that the 'optimal' (hereafter *Crawdad optimal*) clustering has 9 clusters—one for each year. Nearly as good, however, is a three-group clustering in which the 10-Qs are clustered as follows: (1994, quarters 1, 2, 3; 1995, quarters 1, 2, 3; 1996, quarters 1, 2, 3; 1997, quarter 1), (1997, quarters 2, 3; 1998, quarters 1, 2, 3; 1999, quarters 1, 2, 3; 2000, quarters 1, 2, 3), and (2001, quarters 1, 2, 3; 2002, quarters 1, 2, 3). (There are only 3 10-Qs filed each year. The fourth quarter is covered by the 10-K.) Crawdad's optimal clustering of Altria's 10-Ks has four clusters, years 1994–5, 1996–7, 1998–2000, and 2001–2. For Lorillard's 10-Qs, the Crawdad optimal clustering has 10 clusters, but the 3 cluster is given as nearly as good: (1994; 1995; 1996, quarters 1, 2), (1996, quarter 1; 1997; 1998; 1999), and (2000; 2001; 2002). For the 10-Ks, the Crawdad optimal clustering has 3 clusters: (1994; 1995; 1996), (1997; 1998; 1999), and (2000; 2001, 2002). For RJ Reynolds, Crawdad's optimal clustering of the 10-Qs has five clusters: (1994; 1995; 1996), (1997; 1998; 1999), (2000), (2001), (2002). With three clusters (nearly as good) we get: (1994–1999), (2000; 2001), and (2002). Crawdad's optimal clustering of the 10-Ks has four clusters. Nearly as good is a three-cluster arrangement yielding: (1994; 1995; 1996), (1997; 1998; 1999; 2000), and (2001; 2002). The Crawdad optimal four cluster arrangement is the same as this three-cluster one, but with (1997; 1998; 1999; 2000) split into (1997; 1998) and (1999; 2000). These results are broadly

in agreement with the results we reported above with our categorized text query cubes (CTQCs).

Having demonstrated plausibly that text analytics methods can indeed detect linguistic usage patterns in SEC documents (at least for those of the tobacco companies in the 1990s), we now turn to the question of whether such information could be used in evidence.

5. Q3: TEXT ANALYTICS AND THE RULES OF EVIDENCE

Our remarks in this section pertain to American law. We do not claim legal expertise beyond it.

5.1 The Intersection of the Rules of Evidence and Text Analytics

The Federal Rules of Evidence (FRE) establish what may be offered into evidence in court and considered by a jury. The FRE are promulgated by Congress with the advice of the Supreme Court and must be followed in federal courts throughout the United States.[6] By and large, the FRE are liberal and designed to allow, rather than inhibit, the introduction of evidence into court. The foremost consideration is whether something is relevant, which the FRE defines as, "evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence."[7]

The question becomes whether documents are admissible where in the aggregate they are clearly relevant but individually might not be. For instance, one internal email from a supervisor to a worker that contains a sexual joke would likely not meet the threshold for establishing sexual harassment liability. The email might not even be relevant if, for example, it can be proven that the recipient did not even read the email or see the joke. However, ten emails containing sexual jokes would establish a pattern of harassing behavior; the emails in the aggregate are clearly relevant.

---

[6]Individual states are not required to follow the FRE; instead, each state has promulgated its own rules of evidence, most of which largely mimic the FRE.
[7]FRE 401.

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	0.339	0.380	0.275	0.422	0.643	1.015	0.667	0.653	0.686
Altria10Qs:	0.501	0.398	0.258	0.308	0.512	0.587	0.292	0.484	0.277
Lorillard10Ks:	0.178	0.066	0.141	0.185	0.465	0.751	0.410	0.299	0.242
Lorillard10Qs:	0.547	0.404	0.210	0.345	0.620	0.636	0.610	0.445	0.483
RJReynolds10Ks:	0.095	0.201	0.188	0.212	0.502	0.414	0.514	0.484	1.081
RJReynolds10Qs:	0.032	0.178	0.251	0.211	0.314	0.488	0.116	0.567	0.412

Table 5: Normalized Scores. Query: C:social responsibility

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	47.18	39.71	31.47	42.53	34.40	35.94	31.73	28.59	25.35
Altria10Qs:	20.12	22.92	21.21	21.39	15.81	26.66	26.24	25.38	24.99
Lorillard10Ks:	6.110	5.491	16.24	17.47	21.18	21.57	22.16	21.32	19.23
Lorillard10Qs:	13.62	15.67	15.38	22.55	16.43	26.83	30.31	24.92	21.71
RJReynolds10Ks:	33.15	30.76	29.70	26.53	26.45	26.21	16.90	17.81	19.04
RJReynolds10Qs:	16.63	9.785	22.81	24.50	17.43	23.76	16.60	15.28	17.43

Table 6: Normalized Scores. Query: C:profitability

Text analytics would be able to quickly and efficiently search for these documents in order to attempt to find a pattern of harassing behavior. Courts interpret the relevancy rule liberally; something will be deemed relevant often if there is even a slight possibility that it might aid the jury in the determination of guilt, innocence, or liability. It is thus likely that a series of documents that are relevant in the aggregate will be admitted into evidence in toto even if some of the individual documents are not wholly relevant; courts are steadily moving away from strict formalism and towards pragmatic interpretations of the rules.

Once something is deemed relevant, it is presumed admissible unless some exception bars its admission.[8] There are two prominent exceptions to admitting evidence: where (1) the evidence would be unfairly prejudicial to the other party; and (2) where the evidence constitutes hearsay.[9] This first exception is commonly referred to by its rule number, 403, and bars evidence that is admittedly relevant, but "its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury."[10] Put another way, the evidence is relevant, but must not be shown to the jury because it would likely prejudice the other ("nonoffering") party. While this rule is invoked often by nonoffering parties, courts rarely sustain the objection. Nevertheless, a 403 objection is more likely to be sustained when the evidence offered has the possibility of confusing the issues or having an "undue tendency to suggest decision on an improper basis."[11] For instance, in a car accident lawsuit the fact that the defendant once kicked a puppy would be excluded under FRE 403 because it might result in a finding of liability at least partly because the jury dislikes the puppy-kicking defendant.

The second exception constitutes one of the largest parts of the FRE. Hearsay is defined as, "a statement, other than one made by the declarant while testifying at the trial or hearing, offered in evidence to prove the truth of the matter asserted."[12] This is seen by practitioners as clunk phrasing, and lawyers typically restate the hearsay rule as: an out of court statement used for the truth of the matter asserted. Put another way, hearsay is any statement made by someone who was not under oath at the time he said it, excluding instances where a witness in court merely repeats what someone else said outside of court. This is best illustrated by way of example: suppose there is a car accident between Adam and Ben, which Cindy witnesses. Cindy tells the police officer who reports to the scene that she saw Ben run the red light and hit Adam. Adam later sues Ben and calls the police officer to the stand to testify about what Cindy told him. Cindy's statement is hearsay and the police officer would be prevented from repeating what Cindy told him. However, Cindy, if called to the stand, would be allowed to testify about what she saw, since it was a personal observation rather than a repeated statement by another. Essentially, the FRE favors firsthand evidence over secondhand accounts or observations: the rules generally disfavor Witness B testifying about things he heard Witness A say; Witness A should be on the stand testifying about what he saw.

Nevertheless, there are exceptions to the hearsay rule, and the one most relevant for our discussion is that statements made outside of court by a party to the lawsuit are in fact deemed non-hearsay.[13] Using the above car-crash example, suppose that after the collision Ben got out of his car and told Adam, "I'm so sorry! I ran the light–it's all my fault!" Adam then sues Ben, and Adam attempts to testify regarding Ben's admission. This is permissible because Ben is a party to the lawsuit; his statement, although made outside of court and repeated by Adam in court, is deemed under the FRE to be non-hearsay.

In practical terms, during the phase of a lawsuit called Discovery, lawyers from both sides review each other's documents and attempt to ferret out relevant evidence for use at trial.[14] With the growing use of electronic data, more and

---

[8]See FRE 402.

[9]See FRE 403; FRE 801. Thee are additional considerations and exceptions, but Rule 403 and the hearsay are most relevant to the discussion in this article.

[10]FRE 403.

[11]See FRE 403, Advisory Committee's Note.

---

[12]FRE 801(c).

[13]See FRE 801(d)(2).

[14]It should be noted that documents obtained outside of Dis-

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	3.054	3.045	4.128	6.196	9.067	9.610	8.979	7.842	8.881
Altria10Qs:	0.713	1.595	1.834	4.094	3.540	5.944	4.410	5.021	4.243
Lorillard10Ks:	0.729	0.632	1.915	2.817	3.035	2.722	2.812	1.947	1.485
Lorillard10Qs:	2.784	2.262	2.344	3.718	2.374	4.780	2.854	1.730	1.711
RJReynolds10Ks:	2.484	2.082	1.758	5.424	6.030	4.272	4.536	3.634	3.081
RJReynolds10Qs:	0.360	0.178	0.663	0.819	1.050	1.144	1.967	3.176	1.916

Table 7: Normalized Scores. Query: C:health and safety

	1994	1995	1996	1997	1998	1999	2000	2001	2002
Altria10Ks:	8.995	11.80	11.19	13.09	19.16	9.376	9.889	9.585	8.397
Altria10Qs:	4.464	4.571	5.374	11.99	5.321	7.010	6.045	6.752	6.127
Lorillard10Ks:	2.268	1.639	4.871	5.505	14.38	6.177	4.414	5.076	4.992
Lorillard10Qs:	9.448	11.03	8.023	13.26	7.368	6.278	6.777	6.956	6.694
RJReynolds10Ks:	6.593	7.253	9.294	9.147	8.543	6.637	7.742	9.329	7.605
RJReynolds10Qs:	7.040	9.483	6.268	10.28	5.564	6.834	10.04	11.09	8.531

Table 8: Normalized Scores. Query: C:legal

more companies are storing millions of documents on hard drives. When they turn over these hard drives to the opposing counsel during Discovery, it becomes a burden trying to find the needle in the haystack of millions of memos, emails, and reports. Text analytic software helps to sift through the haystack and try to pull out only the documents that are relevant, which, as discussed above, is the first step in determining whether something is admissible in court.[15] If a lawsuit concerns Securities and Exchange compliance, lawyers would want to use software to pull out only documents relating to compliance in some way; a random email about a secretary's new puppy would not make the cut. Once these relevant documents are found, lawyers must apply the exceptions to the general rule that relevant evidence is admissible, including ascertaining whether something is hearsay. If it is an internal document or an email where the sender works for the company, then such documents would be deemed non-hearsay because they are statements made by a party to the suit.

5.2 Finding Patterns

One of the benefits to using text analytics on Discovery documents is that the software can find patterns over time that a person reviewing the documents individually might miss. Suppose internal memos show a CEO discussing his company's use of A in its products. There is then a court ruling making use of A illegal. Thereafter, the CEO ceases discussing A in all internal memos, and instead begins discussing B, which is similar to A but not identical (and not illegal). Such a clear demarcation is useful information: it might suggest that the company is trying to comply with the change in the law, or it might suggest that the company is trying to get around complying with the change in the law. Either way, this information is very valuable to prosecuting or defending the company's actions in court, but

such a change might not be noticeable to the human eye if one were to manually go through the internal memos; the analyst might miss the forest for the trees.

Under the rules of evidence, a change in the CEO's discussion of something in internal memos meets the test for relevancy because such a change would make it more or less likely to prove the company's liability. The memos would also meet the hearsay test: because the CEO is an agent of the company, he speaks on its behalf, and therefore the memos are party admissions. It is also unlikely that the memos would be deemed unfairly prejudicial if admitted into evidence, unless they lay out the CEO's desire to kick puppies or otherwise pose a danger of causing the jury to decide the issue based on emotion. Nevertheless, the law has yet to establish a rule for which there are not a myriad exceptions, and so we come to the next exceptions to admissibility: viz., character evidence and habit.

Rule 404 states that evidence of a party's character is generally inadmissible for proving action in conformity therewith.[16] Put another way, the FRE will not allow past action to prove current action. For example, suppose A has gotten into six car accidents in the last month. A then hits B with his car. In B's suit against A, the fact that A got into those six accidents in the last month is inadmissible character evidence; the fact that A got into those accidents is irrelevant to whether he got into this one. In the context of text analytics, the software might pull out six reports showing Company A documented worker injuries caused by one particular worker, Accidental Dan. Such evidence would likely be inadmissible character evidence in a new lawsuit where an injured worker claims that Accidental Dan caused his injury in the same way as in the prior six incidents. However, an exception to this exception is if Accidental Dan's actions can be deemed habit or routine.

Rule 406 allows evidence of a person's routine practice or habit to show conformity on a particular occasion.[17] Six car accidents in the last month merely show a pattern of bad driving, and are inadmissible character evidence. However, if each accident occurred at one intersection where the driver

---

covery still may be admitted into evidence; something is relevant no matter from whence it came.

[15]Text analytics are becoming an ever greater part of litigation, particularly at large firms. Recently, Morgan, Lewis & Bockius announced it had reviewed one million pages of documents in one month using predictive coding software. See [15].

[16]See FRE 404.

[17]See FRE 406.

routinely ignored the stop sign and drove through without slowing down, then it becomes admissible habit evidence. Using the above example, if the company's accident reports all show that Accidental Dan caused the worker injuries when he negligently secured a particular rig, that would likely meet the test for admissible habit evidence, and the accident reports could come in.

5.3 Expert Testimony

Once a party finds the documents he wants to bring out at trial, the question becomes how to do it. To introduce a document in court, one must call a witness to testify about the document; lawyers cannot simply pull out the document and show it to the jury without first laying a foundation and explaining why the witness is qualified to testify about the document. At trial, anyone may testify if he has personal knowledge of some matter at issue in the case.[18] These are called lay witnesses, and their testimony is limited to either straight facts and observations, or to opinions and inferences that are rationally based on their observations; are helpful to understanding the witness' testimony; and not based on any scientific, technical, or specialized knowledge.[19] This last aspect is the purview of expert witnesses, who may testify on any scientific, technical, or specialized knowledge so long as the expert's expertise is established, and the testimony is based on sufficient facts or data that was the result of reliable principles and methods.[20] Thus, a lay witness is someone who testifies that she saw a neighbor's dog bite a child; an expert witness is someone who testifies that she examined the bite mark and determined it was caused by a Rottweiler.

The test for establishing someone as an expert is very liberal. Someone may be qualified as an expert "by knowledge, skill, experience, training, or education."[21] This can include academic expertise, such as someone with a Ph.D.; it can also include "real life" expertise, such as police officers or, famously, drug users. In U.S. v. Johnson, the Fifth Circuit Court of Appeals ruled that a drug dealer was qualified to give expert testimony about the source and origination of marijuana in the defendant's possession.[22] The court found that the drug dealer's "substantial experience in dealing with marijuana included identification of Colombian marijuana," and it was therefore proper for him to testify as an expert on the matter.[23]

Because text analytic software is not something that most people are familiar with, it would be appropriate under the rules to call an expert to explain to the jury how the software works and why it is useful. Someone with experience using the software would suffice as an expert. However, a final hurdle is establishing that text analytic software is an appropriate subject for expert testimony. This concerns the reliability of the proposed testimony; courts are wary about possibly misleading the jury with pseudoscience that sounds reliable but is not. Originally, the standard was set forth in a famous case called Frye v. U.S., where the Court of Appeals for the District of Columbia held that expert opinion testimony must be generally accepted in the scientific community.[24] However, this proved too restrictive as it foreclosed expert testimony on new and upcoming science, and the Supreme Court later revised the standard in Daubert v. Merrell Dow by establishing a series of factors that federal judges must consider when weighing the reliability of expert testimony.[25] Thus, federal judges must now determine (1) whether the basis of the testimony is scientific (including whether it can be tested; if it has been peer reviewed; what the rate of error is; and whether it is generally accepted in the scientific community); and (2) whether the testimony would help the jury to understand the facts.[26] No one factor is determinative, and as a result the Daubert standard has expanded the scope of permissive expert testimony. With respect to text analytics, the methodology has been around for long enough that it is generally accepted in the scientific community, and therefore so long as the expert is able to explain how the software works, what its rate of error is, etc., a judge will deem it sufficiently reliable for expert testimony.

6. DISCUSSION

To summarize: That we are able, in this exploratory study, to find patterns of language use in the SEC documents has to be counted as very favorable for the prospect of using text analytics for event studies. It is also positive news that our findings with the SEC documents are broadly consonant with the findings of O&M for the CEO's speeches. We should not expect the two sources to be in complete agreement, since the relevant texts have different purposes. CEO Bible's was to frame how his audience should think about Philip Morris and to put forth the best case for the company. The SEC filings are constrained by broadly legal requirements (laws, regulations, etc.) to report certain material facts. Finally, by way of summary, we have also argued that text analytics results of good quality will be found acceptable as evidence, at least in American law.

The situation, then, is most encouraging for continued research and development of text analytics for, broadly, event studies in legal contexts. Moving forward requires recognizing opportunities for future research as well as limitations in present knowledge. Regarding the latter, this has only been an exploratory study. A main worry about the results is that their significance is only apparent because we have somehow "cherry picked" our queries. Surely it is true that if enough queries are posed to the documents, some will be found statistically significant in the (bogus) patterns they reveal. Countering this worry, we note that our queries were motivated in every case by independent findings reported in the O&M study, which findings were for a completely different set of documents, using CRA instead of our concept representation and matching methods. Note further that O&M only dealt with Philip Morris documents, while

---

[18]See FRE 601; 602.

[19]FRE 701.

[20]FRE 702.

[21]FRE 702.

[22]See U.S. v. Johnson, 575 F.2d 1347 (5th Cir. 1978).

[23]U.S. v. Johnson, 575 F.2d 1347, 1361 (5th Cir. 1978). The court was careful to note, however, that qualifying a witness as an expert should not be determinative as to the subject of his testimony. The opposing party is entitled to counter with his own expert, and the jury ought to give whatever weight to the experts' testimonies as they think it deserves. Id.

[24]See Frye v. U.S., 293 F. 1013, 1014 (D.C. Cir. 1923).

[25]Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

[26]See Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 593 (1993).

we have handled the three large publicly held companies, including Philip Morris. Our query representations were formed from a principled borrowing from the O&M findings, augmented by hypernyms and hyponyms from WordNet. We did *not* test and refine the queries, tweaking them to produce statistically significant results. (Of course not!) That there is so much agreement arising from the two approaches is, we submit, quite remarkable. Further, while one in twenty queries in a "fishing expedition" might be expected to be significant at the 5% level, some of our results are significant at the 1 in 13,000 level. This is not something to be expected from a casual, small-scale fishing trip.

Looking forward, there is clearly much to do. More extensive testing is needed in different domains, with different document sets, and on different topics. In addition, we note three specific areas that command near-term attention.

First, standardizing and carefully articulating procedures and methodologies for concept representation and scoring of documents is very high on our list. Informally, as befits an exploratory study, we first identified a concept of interest and assigned terms to it based on the findings of the O&M study. We then expanded the term list using ordinary knowledge as well as synonym sets, hypernyms, and hyponyms drawn from WordNet. The result is a semantically transparent concept representation, which may be shown to and critiqued by non-experts and by other researchers. This is a procedure we believe is well-suited to be standardized for replicability. Further, we expect it will be augmentable by computational procedures, such as classification trees. These are all issues for future research.

Second, broadly statistical means (including exploratory data analytic methods) need to be investigated regarding suitability for these purposes. CRA is evidently very promising as an exploratory tool. The O&M study used factor analysis to augment the data produced by CRA and to identify key themes in the documents. We relied on these results implicitly. Explicit use of factor analysis (and related methods) is also likely to be useful in future studies. As for hypothesis testing, while regression methods are of course on the table and bear examining, our sense (only) is that nonparametric methods, such as the sign test discussed here, are very likely to play a large rôle when these issues are sorted out.

Third, an example of an hypothesis that we have not fully ruled out is that the statistically significant changes we found in the SEC filings are indicative not of the tobacco industry itself, but of general secular trends. Perhaps, for example, it simply became fashionable in SEC filings at the turn of the century to discuss social responsibility. That a similar trend was found by O&M in the speeches of CEO Bible supports our suggested interpretation. In any event, such worries arise with any empirical hypothesis, regardless of methods and techniques used to support it. Credible alternatives simply need to be checked. In the present case one could draw a sample of SEC filings from other industries and determine whether similar statistical patterns are present. However that exercise turns out, the fact remains that text analytics can find interesting and significant patterns bearing on our assessment of events.

7. REFERENCES

[1] J. Agrawal and W. A. Kamakura. The economic worth of celebrity endorsers: An event study analysis. *The Journal of Marketing*, 59(3):56–62, July 1995.

[2] S. Bhagat and R. Romano. Event studies and the law: Part I – technique and corporate litigation. *American Law and Economics Review*, 4(1):141–168, 2002.

[3] S. Bhagat and R. Romano. Event studies and the law: Part II – empirical studies of corporate law. *American Law and Economics Review*, 4(2):380–423, 2002.

[4] U. Bhattacharya, H. Daouk, B. Jorgenson, and C.-H. Kehr. When an event is not an event: the curious case of an emerging market. *Journal of Financial Economics*, 55(1):69–101, January 2000.

[5] S. J. Brown and J. B. Warner. Using daily stock returns: The case of event studies. *Journal of Financial Economics*, 14(1):3–31, March 1985.

[6] S. R. Corman, T. Kuhn, R. D. McPhee, and K. J. Dooley. Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research*, 28(2):157–206, April 2002.

[7] Crawdad Technologies, LLC. Centering resonance analysis: A superior data mining algorithm for textual data streams. STTR Phase 1 Final Technical Report, 24 March 2004. Air Force Office of Scientific Research contract number F49620-03-C0082. Crawdad Technologies, LLC, 5412 W. Harrison Ct., Chandler, AZ 85226-1935.

[8] K. J. Dooley and S. R. Corman. The dynamics of electronic media coverage. In B. S. Greenberg, editor, *Communication and Terrorism. Public and media response to 9/11*, pages 121–135. Hampton Press, Cresskill, NJ, 2002.

[9] B. E. Eckbo, editor. *Handbook of Corporate Finance, Volume 1: Empirical Corporate Finance*. North-Holland, Ambsterdam, The Netherlands, first edition, 2007.

[10] E. F. Fama, L. Fisher, M. C. Jensen, and R. Roll. The adjustment of stock prices to new information. *International Economic Review*, 10(1):1–21, 1969.

[11] A. Keown and J. M. Pinkerton. Merger announcements and insider trading activity: An empirical investigation. *The Journal of Finance*, 36(4), 1981.

[12] L. Marais and K. Schipper. Applications of event study methods in litigation support. In P. B. Frank, M. J. Wagner, and R. L. Weil, editors, *Litigation Services Handbook*, pages 45·1–45·24. John Wiley & Sons, New York, NY, second edition, 1995.

[13] M. d. Oliveira and P. Murphy. The leader as the face of a crisis: Philip Morris' CEO's speeches during the 1990s. *Journal of Public Relations Research*, 21(4):361–80, 2009.

[14] Princeton University. About WordNet. http://wordnet.princeton.edu, 2010.

[15] D. C. Weiss. Predictive coding software helps Morgan Lewis review millions of pages in 1 month. *ABA Journal*, 29 November 2010. http://www.abajournal.com/news/article/ predictive_coding_software_helps_morgan_lewis_ review_millions_of_pages_in_1.

[16] R. S. Wood. Tobacco's tipping point: The Master Settlement Agreement as a focusing event. *The Policy Studies Journal*, 34(5):419–436, 2006.