BA870: Topics in Financial & Accounting Analytics

Lecture #8 (Tuesday, April 14, 2022)

Professor Peter Wysocki

Topics: Intro to “Old School” Textual Analysis o
Financial and Accounting Disclosures

Lecture #8 Agenda

- Review of SEC Database for Textual Disclosures
  - Form Types, CIK Identifiers, Notre Dame Database
  - Financial Disclosure Datasets
- Introduction to Textual Analysis and NLP in Finance and Accounting
- Sources of Data for Textual Disclosures
  - Form Types, CIK Identifiers, Notre Dame Database
  - Financial Disclosure Datasets

SEC Edgar Database (www.sec.gov)

- SEC Form Types (U.S. Listed Companies)
  - Includes many international companies that are directly listed or cross-listed in the U.S.
- Form Types for Textual Analysis
  - 10-K (Annual Financial Report)
  - 10-Q (Quarterly Financial Report)
  - 8-K (Material Events)
  - DEF14A (Proxy Statement)

SEC Form 10-K

- The Form 10-K is an annual report, which "provides a comprehensive overview of the company's business and financial condition.
- This form includes specific audited financial statements.
- Form 10-K must be filed within 90 days of the end of the company's fiscal year.
- Although similarly named, the annual report on Form 10-K is distinct from the "annual report to shareholders," which a company must send to its shareholders when it holds an annual meeting to elect directors.

SEC Form 10-Q

- The Form 10-Q includes unaudited financial statements and provides a continuing view of the company's financial position during the year.
- The report must be filed for each of the first three fiscal quarters of the company's fiscal year.

SEC Edgar Database (www.sec.gov)

- CIK Identifiers
  - WRDS download example
- Notre Dame Database
  - University of Notre Dame Software Repository for Accounting and Finance
  - https://sraf.nd.edu

Textual Analysis of Financial and Accounting Data

- To this point in BA870, we have examined data sources and analytics tools for analyzing **numbers** found in companies' financial statements.

- But, a major part of the information disclosed by companies is textual information (often viewed as qualitative or unstructured information).

- This lecture introduces data sources and analytics tools for analyzing **text/language** disclosed by companies.

What is Textual Analysis and Natural Language Processing (NLP)?

- Natural Language Processing (NLP) is a broad field encompassing the automated interpretation of human language.

- Within NLP, text analysis is the process of extracting meaning from digital text. This original text data could be in the form of documents, webpages, news articles, email, social media posts, etc.

- We will focus on text that might be used by current or potential investors in a company.

Why Use Text Analysis?

- Text analysis is powerful because it can help extract useful information from large quantities of text data.

- Text analysis minimizes the human effort required to consume digital text, and results in quantitative knowledge gathered from the digital text.

Why Use Text Analysis? (cont.)

- For example, if you need to
  - Decide which company to invest in (among thousands of possible companies)
  - Decide when to sell shares of a company (ie, monitor any changes in the company's health or risk)
- Then, you would want to analyze **both** the company's past financial numbers, newly released financial number AND
  - What the company is saying or disclosing about changes its health and its expected future risk and performance

Why Use Text Analysis?

Fraud detection model based on the textual, i.e., content, analysis of MD&A in 10-K:

$$\text{Fraud}_i = 2.89757 - 0.83408\ (\text{Positive Emmotion}_i) - 0.48315\ (\text{Present Tense}_i) + .0001\ (\text{Total Words}_i) - 2.80753(\text{Colons}_i)$$

"Conventional fraud detection measures using ratio analysis and other financial data were either unable to detect the fraud or unable to detect it soon enough to avoid catastrophic outcomes".

Lee, Churyk and Clinton (*Strategic Finance* , 2013, p. 33)

Why Use Text Analysis? (cont.)

- For example:
  - Using one of the tools of text analysis—**sentiment analysis** - you can identify lists of "positive" and "negative" words/phrases in a company's reports (or in the news) and then automatically count whether good things or bad things are happening to the company.

Preparing the Text Data

- Before using more advanced tools such as sentiment analysis, it is useful to learn about some of the required steps used to prepare the data prior to the analysis.

- Even if you are using a vendor's off-the-shelf product for your text analysis needs, learning about the underlying techniques is necessary to consider the assumptions that may have been made during the preparation of the data.

Segmenting Text into Words

- The exact steps for cleaning and preparing the raw text data will depend upon:
  - the source of the text data,
  - the purpose of the analysis, and
  - your programming tools.

- Processing large amounts of raw text for analysis always entails chopping it into smaller pieces.

- Text is a linear sequence of characters/symbols and white spaces. A “token” is a series of characters treated as meaningful unit. Typically these would be words.

Tokenization

- Word tokenization transforms the text into a series of word tokens. At this stage, certain characters—such as some punctuation—are typically also eliminated.

- For example, the above English sentences would be turned into the following list of tokens:

‘tokenization’, ‘transforms’, ‘the’, ‘text’, ‘into’, ‘a’, ‘series’, ‘of’, ‘word’, ‘tokens’, ‘at’, ‘this’, ‘stage’, ‘certain’, ‘characters’, ‘such’, ‘as’, ‘some’, ‘punctuation’, ‘are’, ‘typically’, ‘also’, ‘eliminated’

- The list of tokens becomes input for further processing.

Lemmatization

- ***Lemmatization*** is the process of reducing words down to their base form, or dictionary form, known as the *lemma*.

- The verb "*to eat*" may appear in the text as "*eat,*" "*ate*," "*eats*," or "*eating*." The base form, "*eat*," is the lemma for the word.

- Generally, lemmatization means removing inflectional endings. However, reducing the word to its lemma may also mean substitution of a synonym based on context.

Lemmatization (cont.)

For example, review the following sequence of tokens before and after lemmatization:

Before lemmatization:

‘a’, ‘background’, ‘in’, ‘statistics’, ‘is’, ‘expected’,

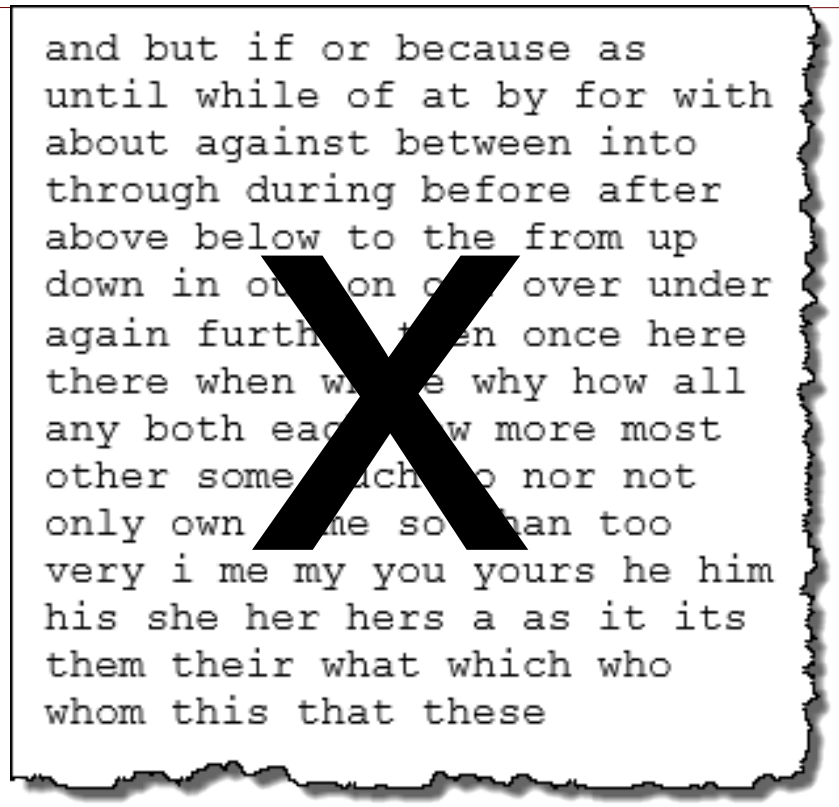
Lemma:

‘a’, ‘background’, ‘in’, ‘statistic’, ‘be’, ‘expect’,

Removing Stop Words

- The next step is typically to filter out the most commonly used words in the English language. These are words used to construct meaningful sentences, but are not very meaningful in themselves.

- These words are called ***stop words***.



There is no universal list.
A typical stop word list has approx. 125-150 words.

Counting Words

- Taking the filtered list and running a word frequency process determines the number of times each token occurs in the digital text being analyzed.

Token	Token Count
made	123
give	122
hope	121
always	119
many	118
us	115
last	115

- *Word frequency*, a measure of the relative weight given to a word in the corpus analyzed, is typically a precursor to further analysis, such as sentiment analysis, readability analysis, and topic modeling.

Adjacent Words

- Collocation refers to words commonly appearing near each other. An example is a bigram, which is two adjacent tokens in a sequence of tokens.

- Returning to the hotel review example, guests often indicate whether they would “stay again” or “not stay” (again). Counting the occurrence of those two bigrams could be relevant as indications of positive or negative sentiment.

unigram = 1 token
bigram = 2 adjacent tokens
trigram= 3 adjacent tokens
n-gram = n adjacent tokens

Example of Tokens and Lemmas

https://wrds-www.wharton.upenn.edu/classroom/natural-language-processing/

This interactive tool uses the entire text of Jane Austen's *Pride and Prejudice* as the source of the text data.

Summary to this point

- Text analysis uses automated processes to extract meaning from digital text.
- Tokenization, part-of-speech tagging, lemmatization, and the elimination of stop words are all techniques used to prepare raw digital text data for analysis.
- Counting the frequency of specific words in the text can provide helpful information about the text, and is a precursor to further analysis.

Textual Attributes That Might Wish to Capture and Quantify

- Similarity between Documents
- Readability Indices
- Sentiment
- Bags of Words
  - Loughran and McDonald

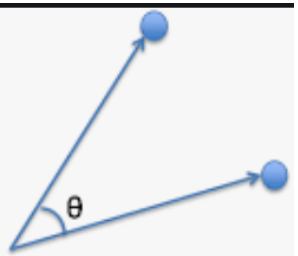
Similarity of documents

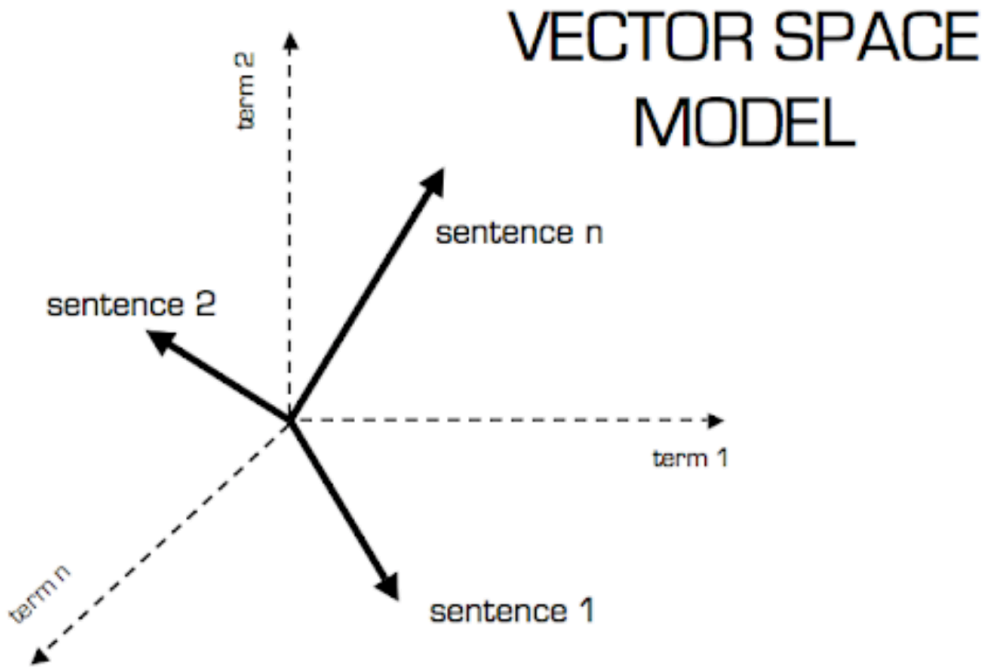
- Similarity between Documents
  - Applications: Plagarism, Removing “Boiler Plate”
  - Measures: Cosine Similarity, Jaccard Similarity

Cosine similarity of 2 documents

- Cosine similarity is a metric used to determine how similar documents are irrespective of their size.
- Mathematically, it measures the *cosine* of the angle between two vectors projected in a multi-dimensional space. The two vectors are arrays containing the word counts of two documents.

$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$





Python code - Cosine

```
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity

a = np.array([1,2,3])
b = np.array([1,1,4])


# manually compute cosine similarity

dot = np.dot(a, b)

norma = np.linalg.norm(a)

normb = np.linalg.norm(b)

cos = dot / (norma * normb)


print(dot, norma, normb, cos,)
```

Readability of a Document

- In Finance and Accounting, there is a concern that company's try hide information form outside parties (or at least not communicate the information clearly).
- How do we measure the transparency or readability of a document?
- Common measures developed in the linguistics literature:

Readability

- Gunning-Fog Index https://en.wikipedia.org/wiki/Gunning_fog_index
- Smog Index https://en.wikipedia.org/wiki/SMOG
- Flesch Reading Ease https://en.wikipedia.org/wiki/Flesch–Kincaid_readability_tests
- Flesch-Kincaid Grade Level https://en.wikipedia.org/wiki/Flesch–Kincaid_readability_tests
- Automated Readability Index https://en.wikipedia.org/wiki/Automated_readability_index
- Coleman-Liau Index https://en.wikipedia.org/wiki/Coleman–Liau_index

Readability of a document

- Readability examples code: See Colab Notebook on QuestromTools.
- Other indices:
  - Bog Index: https://kelley.iu.edu/bpm/activities/bogindex.html

Flesch-Kincaid grade level

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

Gunning fog index

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right)+100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

Automated Readability Index

$$4.71\left(\frac{\text{characters}}{\text{words}}\right)+0.5\left(\frac{\text{words}}{\text{sentences}}\right)-21.43$$

BOSTON UNIVERSITY

Sentiment Analysis of Documents

- Sentiment analysis is the process of extracting and quantifying subjectivity from text.

- Information is extracted from text, regarding:

    - Opinions
    - Attitudes
    - Emotional tone

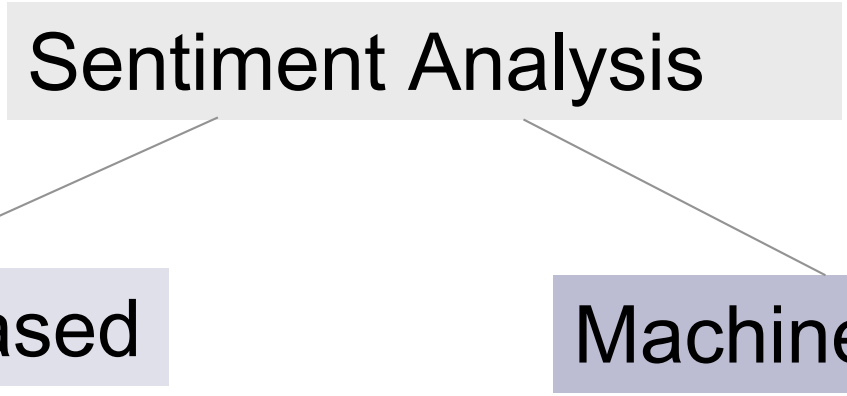
Sentiment Analysis (cont.)

- Sentiment analysis can be done at different levels (e.g., word, phrase, sentence, document, aspect, etc.)
- Sentiment analysis data can be collected in regards to 10-K text, stock market tweets or message posts, analyst reports, conference calls with investors, press releases, etc.

Sentiment Analysis (cont.)

- Sentiment analysis is often conducted by using measures of polarity (positive/negative).
- However, sentiment analysis can go beyond polarity, to classify text according to more complex emotional states, such as happy, angry, or afraid--or particular attitudes, such as litigious or uncertain.

Sentiment Analysis: Two Approaches

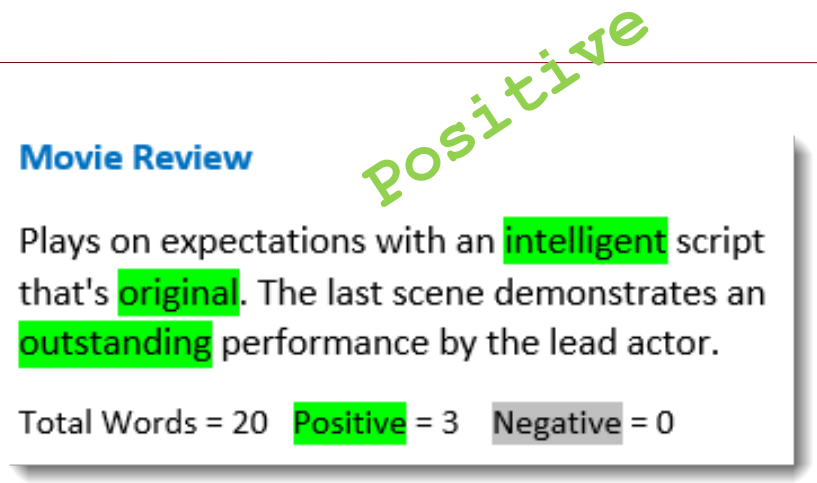


- Uses lexicons – lists of words with labels specifying their sentiments – to identify the sentiment of text.

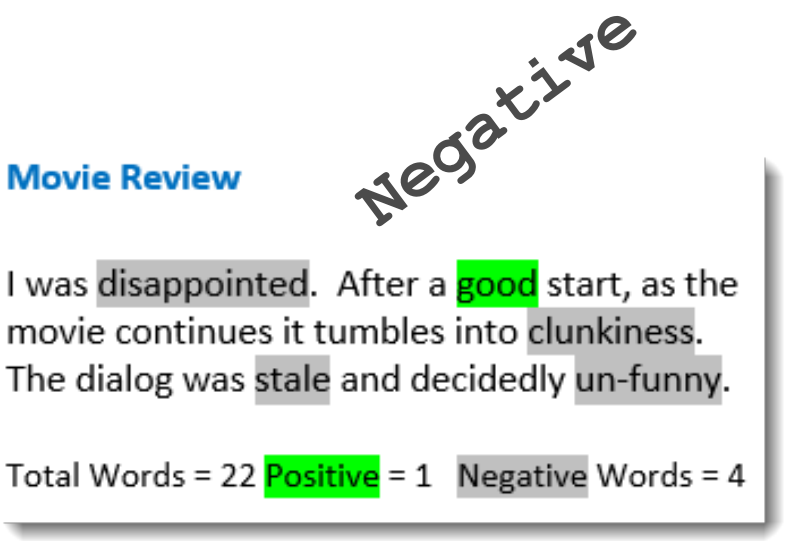
- Uses statistical techniques, enabling computer systems to "learn" with data, without being explicitly programmed.

Lexicon-Based Polarity Classification

- Polarity refers to the position on a scale between positive and negative.



- In a simplistic example of sentiment analysis, text would be classified as positive, negative, or neutral based on how many words or phrases within that text are classified as positive, negative or neutral.



Strength of Polarity

- The polarity classification system typically takes into account a measure of how strongly positive or negative the words or phrases are.
- For example, a scale may be used with +1 being the most positive, - 1 being the most negative, and 0 being neutral.

Phenomenal	+	.97
Incredible	+	.93
Excellent	+	.92
Great	+	.90
Good	+	.89
Okay		.00
Bad	-	.89
Awful	-	.90
Terrible	-	.92
Dreadful	-	.95
Abysmal	-	.97

Sentiment Lexicon Creation

- The creation of the lexicon can be manual (using human annotators), automated, or a combination.

- For example, an automated process may be started by providing a set of seed words with manually assigned sentiment orientations. A computer application can then search for synonyms to expand that seed set.

- Many thousands of words and phrases may be included in the sentiment lexicon.

Machine Learning Approach

- Instead of using a pre-created sentiment lexicon, the machine learning approach analyzes sentiment using statistical algorithms.

- For example, the word “long” would likely be labeled neutral by manual annotators creating a general lexicon. Using a training set comprised solely of movie reviews, a computer program may determine a higher probability that the word “long” appears in movie reviews with a negative rating. In this case, “long” would be labeled negative.

Machine Learning Sentiment Analysis: Early Example

- In 2002, research regarding sentiment analysis using different machine learning techniques was completed using a dataset of online movie reviews.
- In each review, people had already scored the movies using either a star or numerical rating system.
- Based on this rating, reviews were automatically classified as positive, negative or neutral.

ML Sentiment Analysis: Early Example (cont.)

- These classified reviews comprised the “training set” used by the machine learning algorithms.
- Three machine learning techniques were tested: Naive Bayes, Maximum Entropy, and Support Vector Machines.
- Based on the training set information, the algorithms determined the probability of specific words appearing in the text of a positive or negative review.

Lexicon and/or Machine Learning?

The two approaches are often blended, for example:

- Machine learning techniques can be used to create a sentiment lexicon that is then made available for others to use.

- If a training set contains too little data, the machine learning approach may include the use of a pre-existing sentiment lexicon.

Summary to this point

- Sentiment analysis is used to extract opinions, attitudes, and emotional tone from text.
- A lexicon-based approach uses a list of words that have been labeled according to their sentiment.
- In the sentiment lexicon, words or phrases can be classified as positive, negative, or neutral. They are typically given a score relating to the strength of the sentiment.
- A machine learning approach uses algorithms that can "learn" and make predictions regarding sentiment polarity from data using statistical techniques.
- Sentiment analysis can go beyond positive/negative polarity, to ascertain more specific sentiments such as emotions like happy, or attitudes such as litigious.

Conclusion (cont.)

- A machine learning approach uses algorithms that can "learn" and make predictions regarding sentiment polarity from data using statistical techniques.

- Sentiment analysis can go beyond positive/negative polarity, to ascertain more specific sentiments such as emotions like happy, or attitudes such as litigious.

Domain-Specific Language

- Language is context-specific.
- The same terms might have different or even opposite sentiment values in different domains.
- For example, an online video game reviewer may describe a new game as being “sick.” In that particular domain, “sick” may be considered a positive sentiment.

Domain-Specific Language (cont.)

- The informal nature of social media text may contain slang, abbreviations, and alternate spelling variants that do not appear in most dictionaries.

- Even in formal writing, specific topics typically have specialized vocabularies and word usage that would not be represented in a generic dictionary (e.g., medical terminology and law terminology).

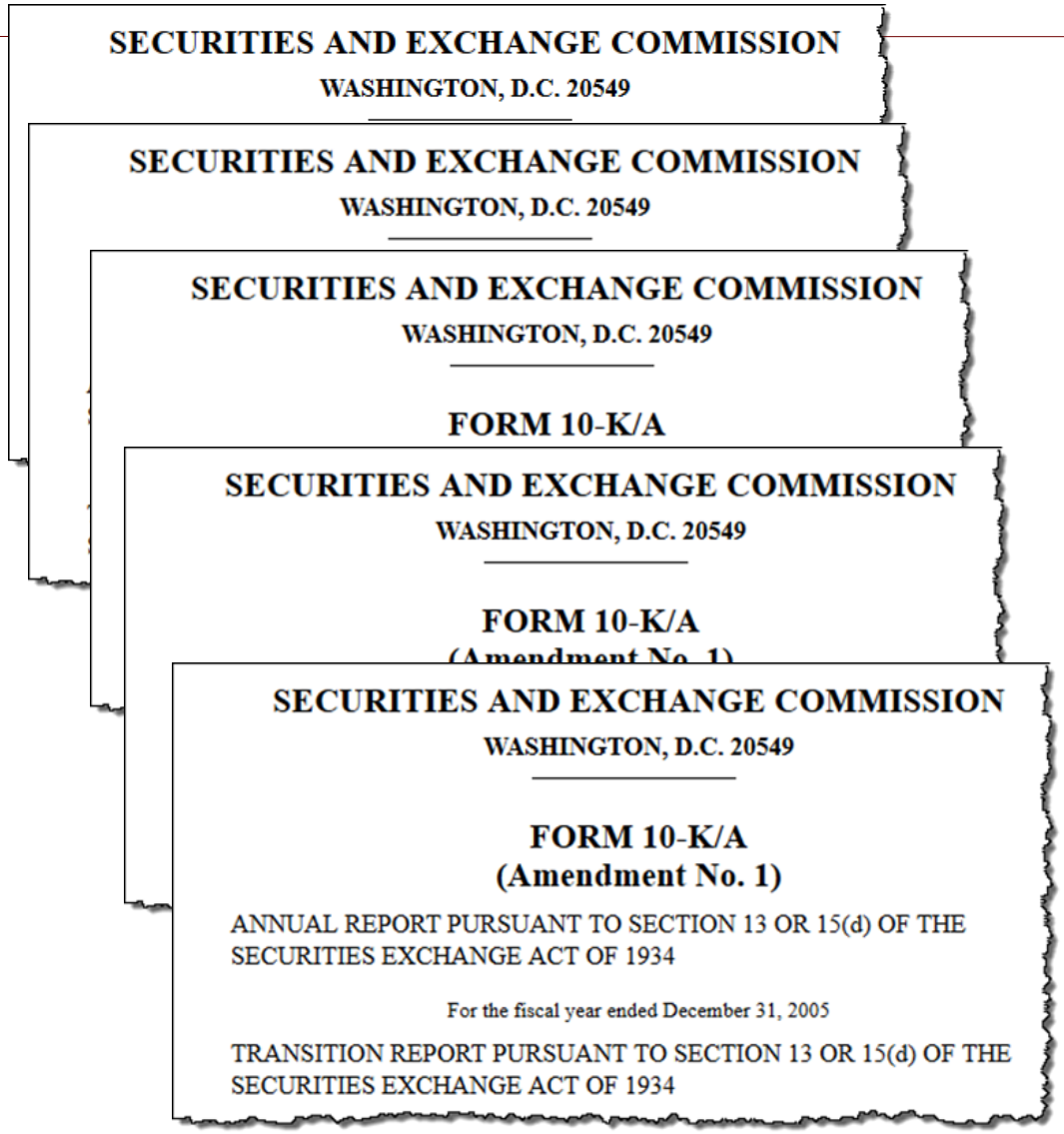
Domain-Specific Language: Loughran & McDonald Study

- In their 2011 paper, Loughran & McDonald discovered how sentiment analysis using a more generic lexicon, developed for human psychology/sociology, could easily misclassify financial text.

- They write, “In the case of finance and accounting, it is best to avoid use of word classifications schemes that come from outside the business domain.”

Loughran & McDonald

- Loughran and McDonald analyzed 41,842 10-Ks from 1994 -2007.
- They examined negative and positive sentiment words from the Harvard IV-4 dictionary in the context of these 10K filings.



Loughran & McDonald (cont.)

- They identified a number of words that were classified as negative in the Harvard IV-4 Psychology lexicon that were typically used in a neutral sense in the context of a company's financial statements. For example:

  - Liability
  - Tax
  - Cost
  - Foreign
  - Depreciation
  - Expense

Loughran & McDonald (cont.)

- In addition to financial-specific words being misclassified as negative, they also found some industry-specific words (e.g., “mine,” “tire,” “crude,” “cancer,” etc.) being misclassified as negative in financial statements.

- They found that almost 3/4 of the negative word counts using the general Harvard list were from words not typically negative in financial or industry-specific contexts.

Loughran & McDonald (cont.)

- Loughran and McDonald also noted some words considered negative in financial documents, such as “misstatement” and “unanticipated,” were not included in the Harvard IV-4 lexicon.

- They created their own six finance-specific sentiment word lists, categorizing words as negative, positive, uncertain, litigious, and either weak modal (e.g., “might,” “could”) or strong modal (e.g., “will,” “can”).

Example

https://wrds-www.wharton.upenn.edu/classroom/sentiment-analysis-lexicons/

- We can use the tool to investigate how selecting different lexicons changes the sentiment analysis of the sample text. The text to be analyzed is the 1999 10-K filed by Coeur Mining, Inc.

Summary to this point

- When choosing a lexicon-based approach to sentiment analysis, it is important to consider how the lexicon was created and whether the lexicon is a good match for your specific sentiment analysis project.

- Sentiment lexicons can be created using various approaches, including manually annotated, dictionary-based, corpus-based, or a hybrid approach that combines methods.

Summary

- Language is context-specific; the same terms might have different or even opposite sentiment values in different domains.

- Tim Loughran and Bill McDonald's 2011 paper concluded that a more general sentiment lexicon was not a good fit when analyzing the sentiment of financial statements; their research showed a domain-specific sentiment lexicon was more accurate.