**Using Categorical Data to Facilitate Learning for Financial Disclosures**

**Applications of the Attribution Dictionary**

Chenhui (Lucy) Li

*Said Business School*

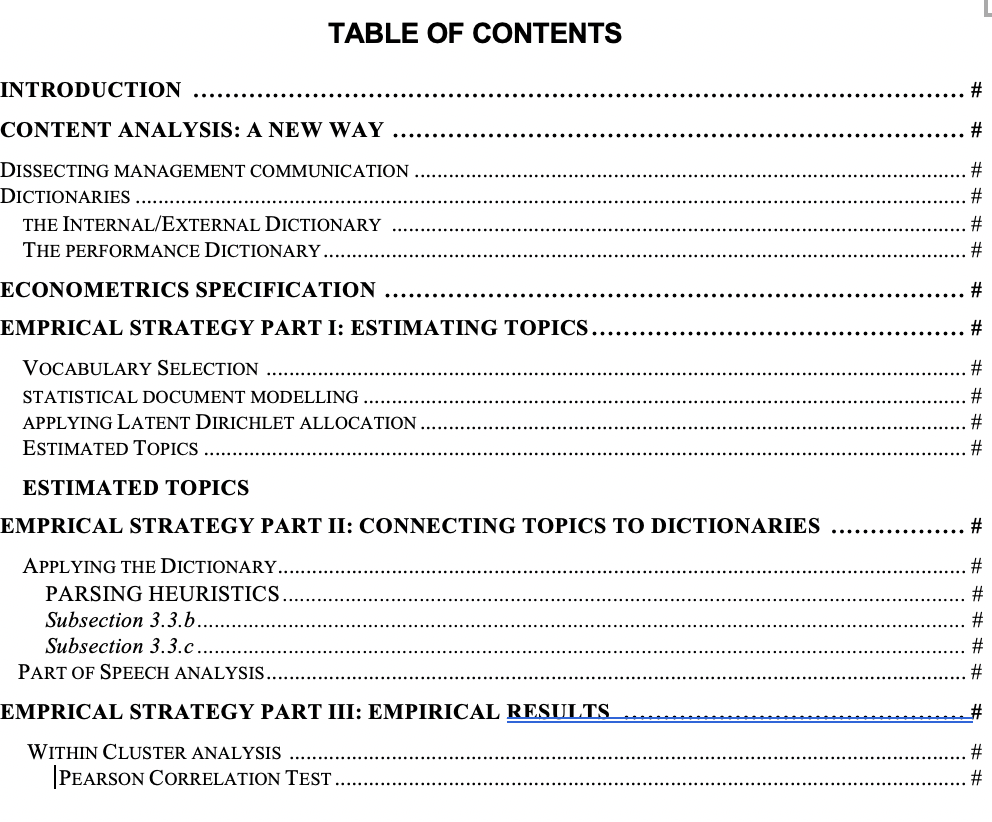
*University of Oxford*

Date

* **Abstract**

This paper discusses the application of the sentiment dictionary which I have jointly developed with Aaron Aujla to analyze the content of management discussion. In the first section of this paper. I demonstrate the relevance of our dictionary with use of topic modelling methods (latent dirichlet allocation and cosine similarities) and study the implications of topic discussion in both a strategic management context and an empirical finance context. I explain the reasons behind my findings with and co-movements of subtopics with fundamental and market based indicators. In the second section of this paper, I explore the joint use of this dictionary with supervised machine learning techniques. Overall, this paper intends a complete analysis workflow focusing on combining the Li-Aujla dictionary with machine learning applications to elicit future research in the area of topic modelling within empirical finance.

# Add implications to management

**WIP**

**1 Introduction**

The efficient market hypothesis stipulates that investors consider all available information in their decision making process (Fama, Fisher, Jensen & Roll,1969; Fama 1970). Financial markets in both quantitative and qualitative forms. Recent studies addressed how financial markets respond to the language in newspaper articles (Tetlock, 2007; Tetlock, Saar-Tsechansky & Macskassy, 2008), earnings reports (Loughran & McDonald, 2011), or other types of regulatory disclosures (Hanley & Hoberg, 2012).

# layout the importance of feature engineering in financial NLP and what the paper offers

The application of natural language processing in finance has become increasingly powerful in financial analysis. Yet, the field remains predominantly focused on analyzing sentiments in financial disclosures, syntactical patterns and word-level characteristics. Analysis on a strategic management or organizational behaviour perspective have mostly been conducted using unsupervised approaches, or approaches that relied on labelled data sets. No research thus far has has looked into

The paper serves to present an approach to integrate this additional field of analysis by integrating topic modelling methods and sentiment analysis. It’s aim is to find the relationship between disclosure on features internal or external to the firm , industry focusing on extending the existing research on sentiment analysis within disclosures with their justifications and industry specific information. For instance, we want to understand questions such as which specific organizational efforts (eg. “strategic acquisition”, “superior management”) has been the most important determinant of success for each cluster of firms ( clusters are formed from firms having similar disclosure characteristics). # more to this

This paper focuses on analyzing MD&A sections of firms’ disclosure. Located in companies’ 10-Ks. MD&A is written to give the investor an opportunity to look at the company through the eyes of management by providing both a short and long-term analysis of the business of the company. It is one of the most read and important components of the financial statements. In 1995, safe harbour provisions of the private securities and litigation reform act encouraged more forward looking information and should make MD&A more informative. On the other hand, the MD&A may not present accurate information because it is not required to be audited. (Hufner, 2007) . Most literature find a significant correlation between current fundamentals and market reactions and textual disclosures.

Past NLP research conducted upon the MD&A section focused on analyzing the linkages between company performance, stock returns and sentiment, time-series consistency and/or the length of corporate disclosure. Some studies have utilized pre-determined dictionaries to assess disclosure characteristics (e.g., uncertainty, tone, competition). Others studies used unsupervised learning methods to associate measures of readability, similarity, deception, or length with firm fundamentals.

#Not sure whether to include econometrics model here

Yet, given that the very purpose of the MD&A is focused on justifying performance with strategy/actions taken by the firm, it is important for a researcher to associate rationales behind performance with the firm's actual performance. I first review how a bag of words approach may be applied to analyzing firm’s strategies. In this section, I explain the usage of a system of dictionaries text classification model I developed jointly with colleagues and its application in a topic modelling context.

I then investigate the use of our dictionary in analyzing the subtopics under each individual topic designated by LDA. We assume that both sentiments and disclosure are related to the topic in the document, and put forward a joint sentiment and topic model, i.e. Sentiment-LDA. After performing LDA on the whole corpus, we assign most likely topics to each document.

Developed in a seminal paper by Blei, Ng and Jordan (2003), LDA is a robust method that relies on statistical correlations among words in a large set of documents to discover and quantify their latent thematic structure (or, topics). LDA is an unsupervised learning model and can be thought of as a dimensionality-reduction technique, similar to cluster analysis or principal components analysis, but designed specifically for use with text. The output of LDA can be used to construct probabilistic measures of the topics discussed in disclosures and their composition of words. LDA allows researchers to explicitly identify and empirically quantify the amount of information analysts discover and interpret in their reports, without referencing the equity market reaction.

I subsequently conduct regression analysis to observe the extracted topics from MD&A to their average stock market reaction. I use the dictionaries to analyze changes in the content of MD&A overtime and its implications on stock market performance. (WIP)

**2 Methods of Analysis**

**2.1 Machine Learning Methods**

* **Literature Review: Supervised Topic Modelling**

Researchers have also been looking for ways to extract the most relevant information from financial press releases or companies’ annual reports and to differentiate between information of different topics. Boudoukh et al. (2013) use an ex ante list of 14 predefined categories (such as “acquisition, deal, legal or award”) to differentiate between relevant news for companies. Their results suggest a strong contemporaneous response of stocks to their media coverage on days which price relevant information arrives but not on days without information. Neuhierl et al. (2013)’s publication investigates the influence of new topics on abnormal returns. They manually classify press releases into major news categories and their subcategories based on content, 10 major news categories, further subdivided into 60 subcategories. They conclude that significant reactions to financial news, news about corporate strategy, customers and partners, products and services, management changes, and legal developments have an impact on share price.

As both supervised machine learning methods and bag of words analysis require annotated data, a lack of annotated domain specific database that is publicly available hinders the development of the technology. Researchers have been making progress on providing academic resources for NLP application of financial news. Malo et al. (2014) trained classifiers to conduct sentence-level semantic analysis for financial news and provided a Financial Phrase Bank consisting of a set of 5,000 sentences, manually annotated by 16 subject experts. This resource was updated by Sinha et al. (2019), who also released an entity-annotated news dataset containing over 12,000 headlines and their related financial sentiment. Oliveira et al. (2016) produced a stock market sentiment lexicon, which includes 20,551 items extracted automatically from microblogs (StockTwits and Twitter)

Gathering labelled data for the task is strenuous and requires the expertise of financial analysts. At present, there is no available tagged dataset for feature engineering. Gooding and Trisboe (2019) used paid data from All Street Research, containing 3097 instances, with categories defined by analysts which they narrowed down to 1824 examples and 11 categories. This study suffered from a small sample size: several categories containing less than 100 examples which meant that they were not enough to train and test.

* **Literature Review: Latent Dirichlet Allocation**

At a document level, topic modelling may also be conducted via latent dirichlet allocation or cosine similarities. Jin et al (2013), for instance, employed latent dirichlet allocation to reduce Bloomberg news articles into 30 topics and manually aligned them with currency fluctuation so that related topics could be identified. Feuerriegel et al (2016). used the common LDA to extract 40 topics from German ad-hoc press releases and discovered that some topics significantly affected abnormal returns of stocks.

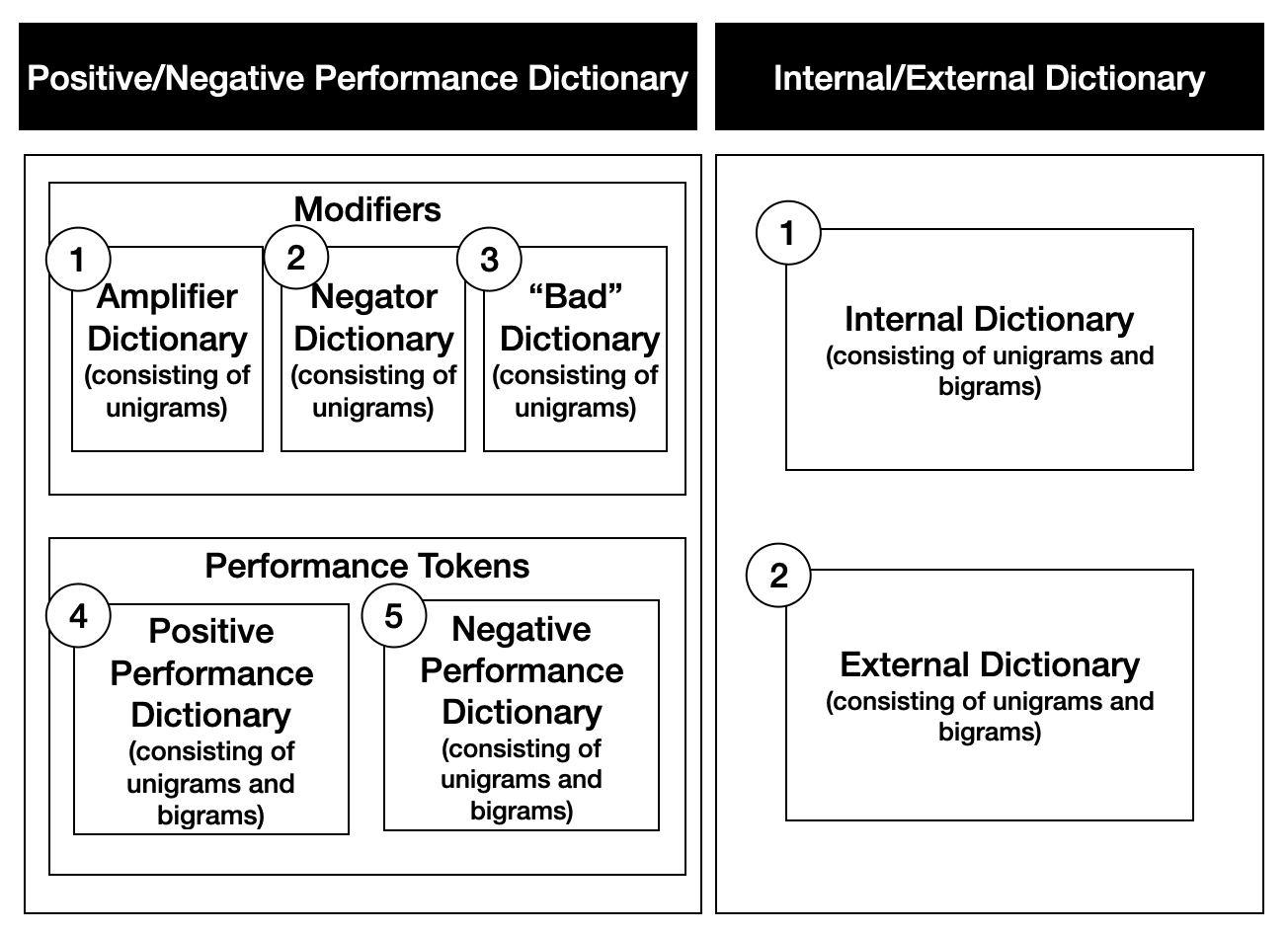
Hoberg, and Maksimovic (2014) use LDA to extract topics from the MD&A part of corporate 10-K filings and to measure corporate disclosure quality based on the topics extracted. Hanley and Hoberg (2016) identified a list of potential systematic risk through LDA and Semantic Vector Analysis (SVA). Bao and Datta (2014) use a variation of the LDA model to summarize the risk-related topics contained in the risk disclosure section (section 1A) of corporate 10-K filings. Curme, Preis, Stanley & Moat (2014) use LDA to identify the semantic topics within the large online text corpus of Wikipedia. The identified topics are then used to determine the link between stock market movements and how frequently Internet users search for the most representative words of each identified topic. Huang, Lehavy, Zang & Zheng (2014) employ LDA topic modeling to compare the thematic content of analyst reports and the text narrative of conference calls. Consistent with the information discovery role of analysts, Huang et al. find that analyst reports issued immediately after conference calls contain exclusive topics that were not dis- cussed during the conference calls. Bao & Datta ((2014) discover and quantify the various topics discussed in textual risk disclosures from annual 10-K filings (Item 1A). The results indicate that about two-thirds of the identified risk topics are uninformative to investors, con- sistent with the notion that risk disclosures are largely boilerplate. Of the remaining topics, disclosures of systematic macroeconomic and liquidity risks have an increasing effect on in- vestors’ risk perceptions, whereas topics related to diversifiable risks (i.e., human resources, regulatory changes, information security, and operation disruption) lead to a decrease in investors’ risk perceptions. Dyer, Lang and Lawrence (2016) use LDA to examine specific topics and find that new FASB and SEC requirements explain most of the increase in length of the topics.

**2.2 Dictionary Methods**

* **The Internal/External and Performance Dictionaries**

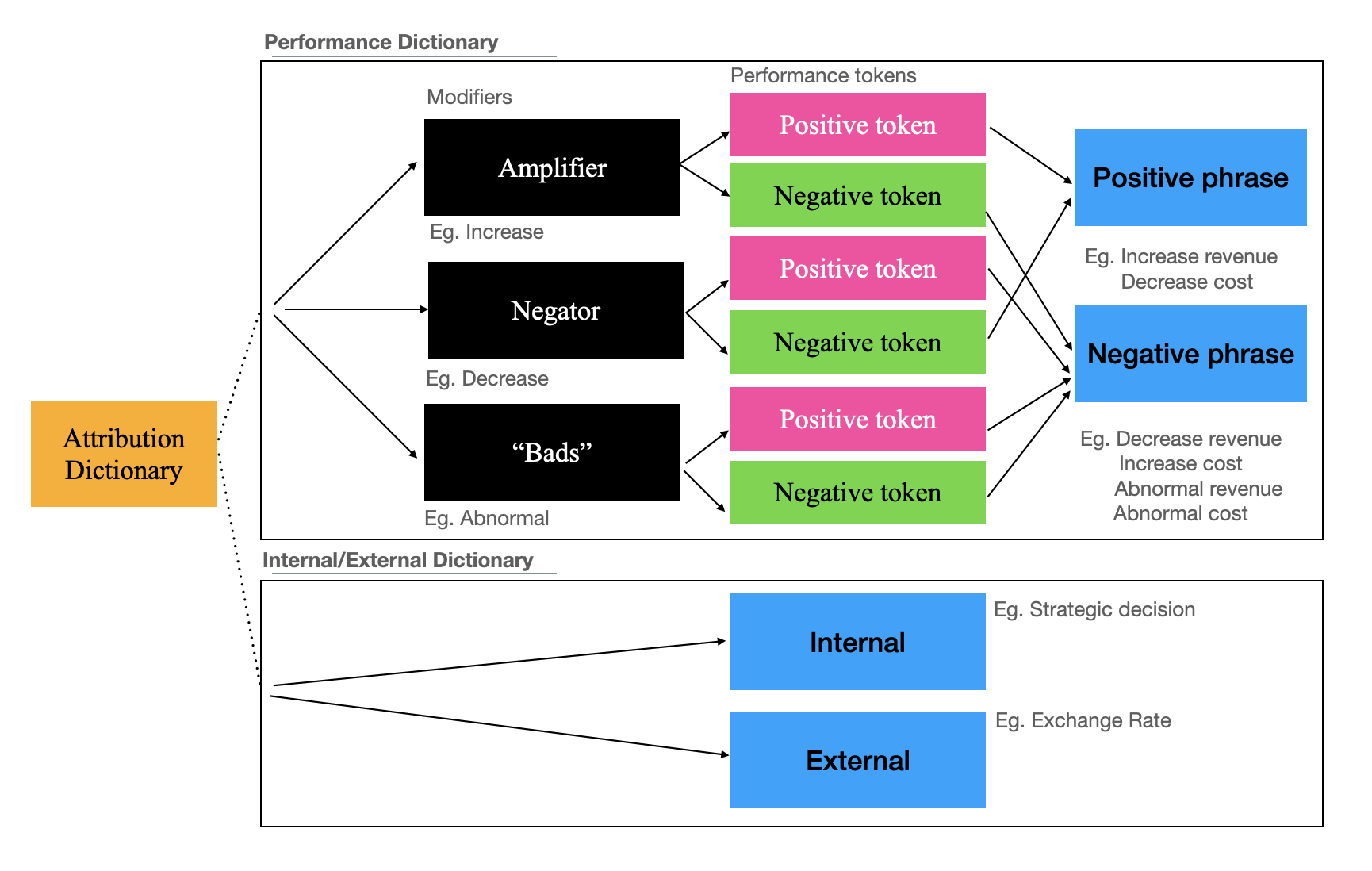
**Fig 1: Dictionary Composition.**

Our positive/negative performance dictionary consists of 5 sub dictionaries and internal/external dictionary consists of 2 sub dictionaries.



**Fig 2: Relationship between individual dictionaries**

Performance phrases are formed by joining amplifier/negator/bads with positive/negative tokens, a performance phrase consist of 2-3 words. Phrases in our internal/external dictionary consist of 1-2 words each.



Our dictionary consists of two parts: an internal/external dictionary featuring key terms justifying business performance, and a performance dictionary consisting of amplifiers, negators, bads, positive and negative performance vocabulary.

An “amplifier” enhances the polarity of the finance specific vocabulary that it is attached to (eg. increased, enhanced, booms, surges, acquire, retain, etc), most amplifiers we have selected are unigrams. A “negator” negates the polarity of the finance specific vocabulary that it is attached to (eg. decreased, reduces, etc). Word in the “bad” category directly makes the polarity of the finance specific vocabulary negative (eg. adverse, constrains, etc).

The positive performance dictionary consists of financial performance tokens and business activities that when amplified, benefits the business (eg. revenue, sales, income, acquisition asset) Conversely, the negative performance dictionary consists of financial performance tokens and business activities that when amplified, negatively affects the business (eg. costs, risks)

The positive and negative performance dictionary consist of financial performance tokens, we have included tokens that have a unique meaning. For instance, “accounting cost” is not incorporated as costs are incorporated into our dictionary and “accounting” does not add an additional layer of meaning to “cost”. For the same reasons, keeping“amortization certain” is not meaningful because “certain” does not add to “amortization”. “Advertising budget” is not included as “advertising” does not add to the polarity of “budget”. However, “Debt maturities” is relevant because both debt and maturity are significant to the polarity of the phrase “Expense reimbursements” is relevant because reimbursement alters the meaning of expense.

Phrases in our **internal dictionary** are defined to fall under any one or more of the following categories.

**Category I: Firm decision making:** the maintaining/altering of management decisions/actions or evaluation of management/managerial intentions, such as investment into research and development, discussion about opportunities, etc.

**Category II**: **Firm competencies:** Innovations, innovative skills, technology, license, intellectual property, rights, rights to patents, copyrights, trademarks, brands, hallmarks, service marks, technical competence, strategy, other forms of abilities, etc. Tokens under this category may align with Barney (1991)’s resource based view, which includes physical capital resources, human capital resources, organizational capital resources, production/maintenance resources, administrative resources, or organizational learning resources and strategic vision resources.

**Category III: Firm actions:** such as M&A activities, partnerships, and investment undertaken, conflict resolution, etc.

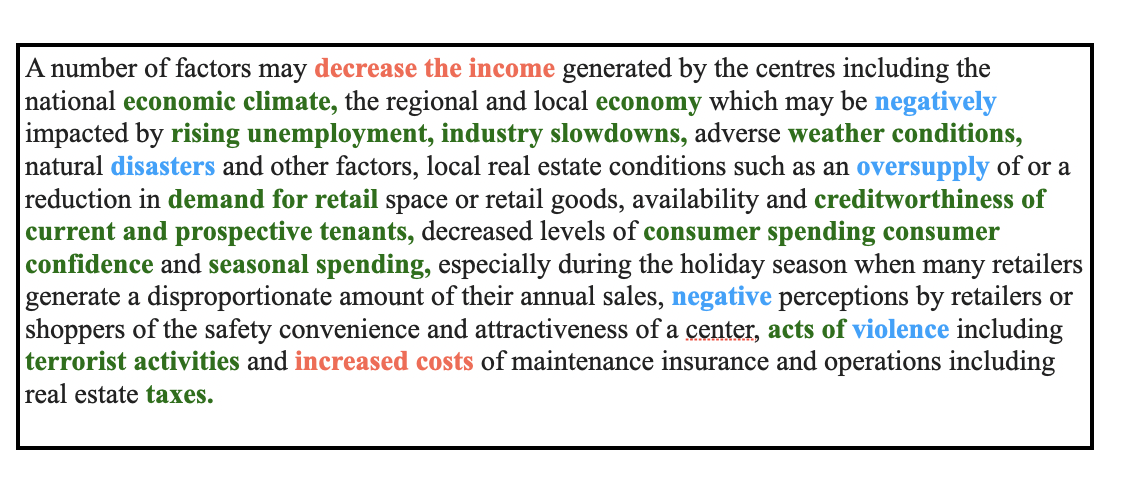
**Category IV: Results/Inference from firm actions:** such asfees/costs revenue originating from firms’ actions (eg. distribution fees, margin ratios). There may be overlaps between tokens that fall under this category and tokens in the performance dictionary, “advertising budget” is an example that is both internal and falls under our positive performance dictionary.

Phrases in our **external dictionary** are defined to fall under any one or more of the following categories.

**Category I: Porters’ 5 Forces:** Analysis of the competition faced by the business, such as competitive rivalry, supplier power, buyer power, threat of substitution and threat of new entry, etc.

**Category II: Institutional or Regulatory Factors and Economic factors:** Geopolitical tensions in recent elections, governmental legislations, lawsuits, exchange rates, taxes, risk, foreign currency, fluctuations, interest rates, foreign currency, forward, option, forward position, option position, other shocks such as natural disasters, adverse weather conditions, cyber security threats, terrorism, etc.

Consider the following paragraph for an example of how our system of classification differs from that of the Loughran and McDonald dictionary.



Whilst the Loughran and McDonald dictionary identifies the blue tokens that either connote positive or negative meaning, our dictionary, in addition, identifies tokens that fall under the performance vocabulary categories (in red) and internal/external categories (in green).

There may be overlaps between the Loughran and McDonald dictionary and our dictionaries. For instance, a sentiment word in the LM dictionary can be referring to a performance token (eg. “negatively affecting revenue”). This will be taken into account as a part of our performance dictionary, however, sentiments that are not associated with the company’s own performance (eg. negative perceptions of retailers and shoppers) is not assessed by the output of our performance dictionary.

**3 Econometrics Specification**

**http://web.sgh.waw.pl/~jmuck/EoPD/Meeting1\_2018.pdf**

Since we want to measure the effects of disclosure content on financial indicators, we would want to conduct analysis on several levels. Our overall econometrics specification estimates the overall effect of disclosure measures on the financial performance (y variable). We use a pooled model where data from different units of our independent variables are pooled together with no assumption on individual differences.

+...+

* E(u)=0
* E(uu’)=0
* rank(X)=K+1<NT
* E(u|X)=0

**4 Empirical Strategy Part I: Estimating Topics**

**4.1 Vocabulary Selection**

**Sample Selection**

I webscrape 10-K filings filed electronically with the SEC on EDGAR from 1993 to 2018. This yields 149,139 10-K and 10-K405 filings from which we were able to parse the Management’s Discussion and Analysis. Management’s Discussion and Analysis comprises Item 7, which usually describes the results of operations, internal and external factors relevant to the business’ performance, and Item 7A, Qualitative and Quantitative disclosures about Market Risks. We have excluded 10-K-A from our sample and eliminated disclosures that contain fewer than 1000 characters, which are disclosures without material information or disclosures that have MD&A section incorporated by reference to the annual report (usually incorporated into exhibit 13, which is kept as a separate file to the main 10-K filing on SEC Edgar). In the latter case, similar to Loughran and McDonald (2011) we find that the beginning and ending positions of the MD&A document when filed in an exhibit are not demarcated in a manner that facilitates accurate parsing. Aside from the MD&A section, exhibit 13 often contains financial statements, notes, as well as the auditor's report, all of which is irrelevant for the purpose of our exercise.

The matching algorithm is designed to capture the position of “item 7. management’s discussion and analysis” and “item 8. consolidated /audited financial statements” and extract the text in between when it satisfies certain heuristics. We remove all numbers and numerical tables, keeping only the text. We identified 87,834 MD&A sections that fulfill this requirement.

I subsequently perform several processing steps that are common in text mining (Manning and Shütze, 1999). We transform the running text into a matrix notation that would allow for further calculations with Scikitlearn’s vectorizer. First of all, we remove stop words that frequently occur in the English language. We use the default NLTK’s list of English stop words which consists of common, short function words that do not add additional meanings to our text - examples of these are conjunctions and pronouns such as “ourselves”, “her”, and “between”.

**4. 2 Vocabulary Modelling**

* **Statistical Document Model**

First consider a document model of the following form:

* Words are represented using unit-based vectors that have a single component equal to 1 and all other components equal to 0. As a result, the vth word in the vocabulary is represented by and for .
* A document is a sequence of N words denoted by )
* A corpus is formed by a collection of M documents: }

**Latent Dirichlet Allocation**

LDA is a generative probabilistic model (Blei, 2012). The underlying assumption of LDA is that it assumes that every document contains a mixture of hidden (latent) topics. Over a distribution of topics, we can infer a topic (and assign a document to it) by choosing the topic that has the highest probability given by a set of words (Blei, Ng, & Jordan, 2003; Blei, 2012).

(i) High Level Summary

LDA assumes that every document will share the same set of topics but exhibit topics in different proportions. A document consists of N word positions, we first populate the word positions with the topic from which the word will come from. We then examine the probability mass function from which the word is drawn, and select a word from the topic we picked.

(ii) Model of LDA

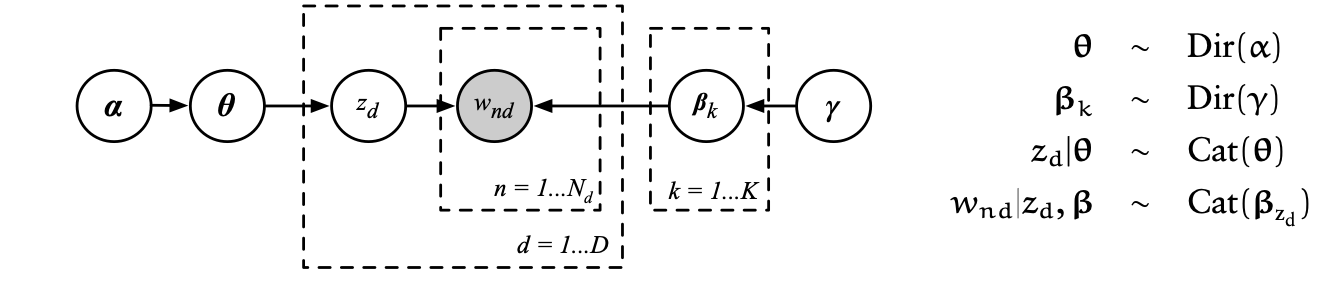


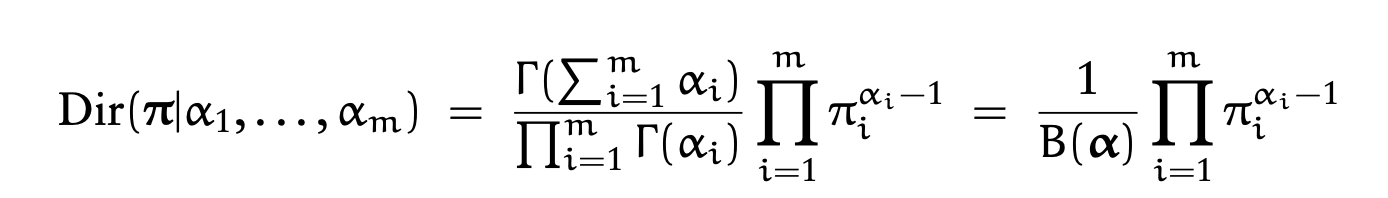
Fig. Description of each stage of the sampling process

* 𝛂 is the proportion parameter.
* nd is the observed variable of words
* d is the per-document topic proportions
* d is the per word topic assignment
* βk is topics
* Gamma is topic distribution

A topic, k, denoted by βk , is a probability mass function over the entire vocabulary. A topic proportion for document d, denoted by d, is a probability function (mixture) over topics for document d.

is a k-dimensional dirichlet variable (where the dimensionality is the predefined number of topics we extract from the documents). lies in the (k-1)-simplex, if i ≥0, .

We say that the density function of ~Dir(𝛂). The Dirichlet distribution is given by:



The dirichlet distribution is used because it is conjugate to the multinomial distribution and has finite dimensional sufficient statistics.

For each topic k ∈ *{1,...,K},* we draw a topic proportion βk from a Dirichlet Distribution Dir(𝛂). , the K-th dimensional probability vector the dirichlet distribution yields, goes into a multinomial distribution. varies with alpha, when alpha is 1, the probability of each outcome is equally likely. As we vary the parameter alpha, we arrive at different points where the multinomial will land.

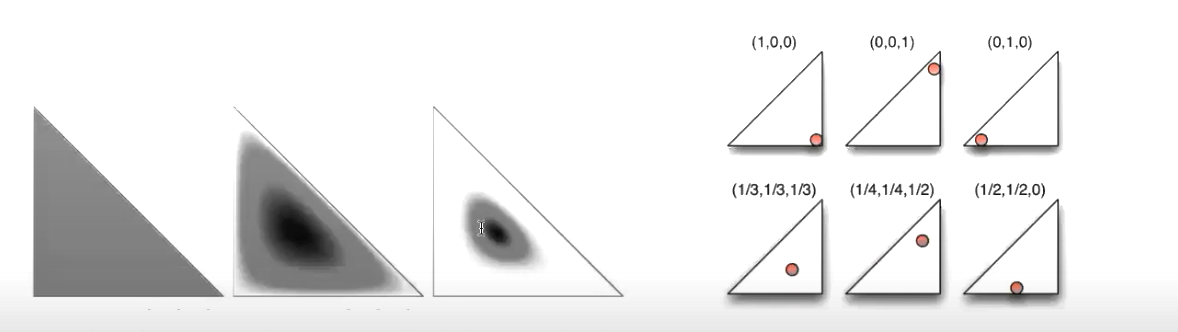
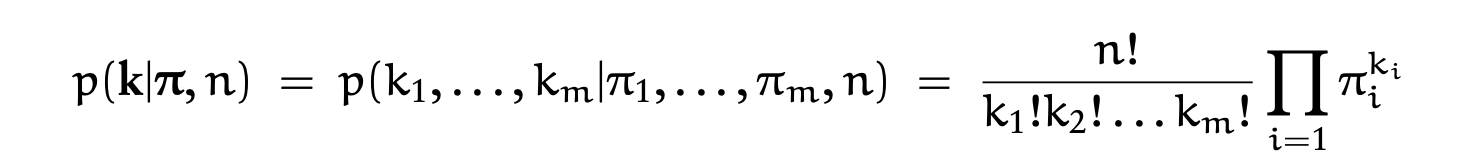


Fig. Dirichlet distribution visualization on a 3-Simplex versus multinomial distributions

The multinomial distribution is a generalization of the binomial function. That is, for n independent trials, each trial is placed into exactly 1 of k categories. The multinomial distribution models the probability of n independent trials each of which leads to a success for exactly one of k categories.

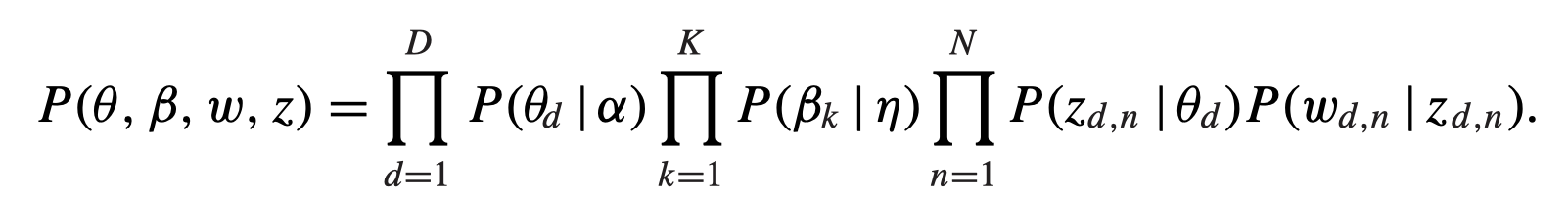


Then, for each document d ∈ *{1,...,M},* we draw a multinomial distribution from a dirichlet distribution with parameter .

Subsequently, for each word position, n ∈ *{1,...,N},* we select a hidden topic zn from the topic proportion for the document using the multinomial distribution from the previous step.

Then, for the word position n, we select a word from the corresponding topic βzn, using the topic selected in the previous step.

Hence, the joint likelihood expression is given by:



**4.3 Construction of Global Topics**

The majority of categories constructed with LDA relate to the sector/industry of the business. It thus may lead to further analysis into the the direction of sentiment attribution in these individual segments. show that the method is suitable for modelling the sector/industry of the business.