VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

MASTER THESIS

The Use of News in Financial Markets:
Evidence from Big Data and Topic Models

Naujienų Nauda Finansų Rinkose:
Įrodymai naudojant didelius duomenis ir temų modelius

Šarūnas Merkliopas

VILNIUS 2016

Ekonometrinės analizės katedra

Darbo vadovas prof. Marijus Radavičius
Darbo recenzentas lekt. dr. Vaidotas Zemlys

Darbas apgintas 2016 m. sausio 14 d.
Darbas įvertintas _____

Registravimo Nr. 111000-9.1-5/_____
2016-01-04 _____

The Use of News in Financial Markets: Evidence from Big Data and Topic Models

Abstract

In this master thesis, the famous Thomson Reuters corpus is analyzed using modern statistical models for text data to investigate a novel research question - just how important the content of news is to financial markets? Using topic models, topic mixtures or proportions of topics that news write about within documents (time periods) are inferred and then used to predict market outcomes or test Granger causality between topics and market variables, where topics are treated to be probability distributions of words. Results of the experiments point to the direction that the content of news might not correspond to the market outcomes exactly and do not carry significant amount of predictive power. However, Granger causality testing reveals, that some specific topics, for example regarding Federal Reserve, politics and etc. do seem to Granger cause market variables and especially, the volatility of trading. This seem to suggest that the content of news might be important as a mean for market players to get informed but other factors are at play too which determine how markets are going to behave. Additionally to this research question, inference algorithm using Direct Representation scheme is derived for Supervised Hierarchical Dirichlet Process mixture model which is found to be superior to some other representation schemes in the literature.

Key words: topic models, text data, financial markets, Reuters corpus, prediction, causality

Naujienų Nauda Finansų Rinkose: Įrodymai naudojant didelius duomenis ir temų modelius

Santrauka

Šiame magistriniame darbe panaudojamas Thomson Reuters tekstų rinkinys kartu su moderniais statistiniais modeliais skirtais analizuoti tekstinius duomenis siekiant ištirti neįprastą klausimą - kiek svarbus finansų rinkoms yra naujienų turinys? Pasitelkiant temų modelius, yra įvertinami temų mišiniai arba temų apie kurias rašo naujienos proporcijos dokumentuose (laiko perioduose), kurie vėliau panaudojami prognozuoti rinkų elgseną arba ištirti Grandžerio priežastingumą tarp tų temų mišinių ir rinkos kintamųjų, kur temos suvokiamos kaip tam tikri tikimybiniai žodžių pasiskirstymai. Eksperimentų rezultatai parodė, kad naujienų turinys tiksliai neatspindi rinkos rezultatų ir neturi ženklios prognostinės galios. Kita vertus, Grandžerio priežąstgumo tyrimas atskleidė, kad kai kurios temos, pavyzdžiui apie Federalinę Rezervų sistemą, politiką ir t.t., sąlygoja rinkos kintamuosius, o ypač - rinkos apyvartą, Grandžerio prasme. Tai sufleruoja, kad naujienų turinys yra svarbi priemonė rinkos dalyviams būti informuotiems, bet rinkų elgseną nulemia ir kiti veiksniai. Darbe taip pat išvedama Tiesioginė Reprezentacijos schema Hierarchinio Dirichlės Proceso mišinių modelio su apmokymu vertinimui, kuri literatūroje nurodoma esanti pranašesnė už kai kurias kitas schemas.

Raktiniai žodžiai: temų modeliai, tekstiniai duomenys, finansų rinkos, Reuters rinkinys, prognozė, priežastingumas

Contents

List of Tables

List of Figures

1 Introduction

Financial markets have long attracted scholars as an object of study due to its complex nature. Currently there are several positions among the researchers regarding the level of predictability of the financial markets and the rationality of market players, but it would be safe to say that at least the problem of forecasting here is difficult and sentiments (in a broader sense - subjectivity) of investors play an important role [16], [11], [20].

Unlike other relevant types of data (e.g. price/volatility data), text data is much more abundant, less structured and characterize sentiments or subjectivity of market participants directly. As a result, new fields of study have emerged such as Sentiments analysis or modeling/predicting financial markets' behavior based on text data ([15], [16], [11]). On the other hand, these approaches somewhat ignore the structure of the texts and focus more on characterizing the relationship between the style of the analyzed texts and financial markets rather than the substance of texts and markets.

As such, potentially interesting and relevant questions have been left underexplored. Thus, the main aim of the paper is to investigate just how important is the structure of financial news or the content of news to market outcomes and whether it could be used for prediction. The idea is to use topic modeling approach [1], that is to treat text data as consisting of many topics which in turn affect the financial markets. It seems to be intuitively reasonable: conditionally on the fact the financial markets are affected by some text data (i.e. news), different content or topics of the data should have different impact on market participants. On the other hand, it might be possible, that not the topics or structure of the news text have the biggest impact on the market outcomes but rather how they are presented - the sentiment that is being expressed in the news.

To achieve the set aim, the literature is surveyed to identify potentially fruitful approaches and various strategies are tested empirically. The contribution of the paper is twofold. First, it applies recently proposed topic modeling techniques to investigate a novel and relevant research question which has not been done before. It also proposes a possible improvement for one of the models which could be useful in many other contexts but is presented in the appendix.

2 Related work

What makes text data special is its abundance and highly unstructured nature. It is hard to run some kind of causal or predictive models on text data because it is not clear what and how is related in the texts. Due to both reasons, only relatively recently models incorporating or analyzing text data got scholar attention. Below works relating text data and finance (or economics more generally) are reviewed.

2.1 Text data and sentiment analysis

As amount of data and text data increased both industry participants and researchers got interested in using this source of information for modeling sentiment expressed in the texts, for example to quantify what kind of sentiment is expressed in the movie reviews which later could be used to analyze the opinions of the viewers or classify the movies accordingly (for a review of recent developments see [15]).

Most works are based on the methods of extracting the so called features from the texts (which could frequency of specific words or part of text occurrences etc.) and then using them to train a classification model which are then presumably is able to generalize on the unseen or new documents. However, this approach does not model the structure of the texts and thus the main point of such exercise is to extract whatever helps the identification of sentiment the most. This in turns means that there is at least some ad-hoc analysis involved.

Sentiments have also seen a lot of academic attention in finance, but for the most part it has been utilized using so called event studies, time series market data or proxy derivatives which hopefully capture this characteristic of investors [12]. As such, text data which should directly (if some model could perfectly extract the signal from text data) reveal this information has seen much less use.

It also is important to differentiate between investor sentiment and textual sentiment as noted in [11]. The difference between the two is that the former captures subjective judgments of investors while the latter can include it as well but also contains information about conditions in the market itself. Interestingly, their causal relationship has not yet been investigated. However, some argue that this qualitative information in text data may provide additional variation which could be used as a more independent test of market efficiency [14].

But how to capture this qualitative information from text data? It turns

out it might be technically simple: [22] not only shows that anxiety and excitement can be differentiated by analyzing specific words in financial text documents but goes on and provides lists of finance-specific words that can do that. So basically by analyzing occurrences of these words in financial text documents we can hope to characterize the textual sentiment behind the documents.

In fact in the upcoming paper [17] these lists are evaluated. The paper finds that not only they could be used to construct a intuitive and visually appealing sentiment measure, but also that the constructed measures correlate with financial market events and serves as a leading indicators of other commonly used financial indicators that are used to proxy for investor sentiment. For example, one of the measure they used was:

$$Sentiment[T] = \frac{|Excitement| - |Anxiety|}{size[T]}$$

That is - textual sentiment expressed in document T was simply characterized by the difference of number of excitement expressing words used and number of anxiety expressing words used in the documents scaled by number of characters in the document.

2.2 Text data cont'd: are sentiments predictive?

In finance context, sentiment are associated with behavioral paradigm which stands in contrast to Efficient Market Hypothesis and assumption of agent rationality.

One notion of irrationality in finance could be described by introducing the concept of 'noise trader' [20]: basically it is a market participant who makes decisions based on trend, does not weight in market timing or fundamentals and thus tends to missreact to good and bad news. In fact same survey notes that noise traders have been identified as a major source of volatility. If so, it can be of potential use in the current paper as in this case text data such as news which presumably is among the sources of information used by noise traders could have some predictive power of market volatility.

Similarly, there has been considerable debate in the recent literature as to whether investor sentiment predicts stock returns, [12] notes that since rational risk-based asset pricing models predict that prices reflect the discounted value of expected future cash flows, irrational investors in the market are offset by arbitrageurs, so investor sentiment does not affect the price

in fundamental way. However, behavioral finance suggests that investor sentiment, as reflected by retail investor demand, may cause prices to deviate from the underlying fundamentals even for extended period of time as a result of actions of noise traders.

Interestingly, the paper using very big dataset of texts from Yahoo! message boards find no evidence for sentiment influencing stock price at firm/aggregate level or on temporal/cross-sectional dimension. In fact it finds the reverse evidence - that stock performance influence investor sentiment. However, it is important to note the data source might not be representative of biggest market participants, that is big financial institutions like banks and hedge funds and some specific stocks are selected rather than market indexes. Similarly, [19] argue that sentiment index is not sufficient to fully characterize the nature of noise traders as other factors (firm-specific information, liquidity, etc.) affect trading behavior too.

Somewhat differently, [13] constructs sentiment indicator from Financial Times news articles and finds that it is highly predictive of interest rate decisions by Federal Open Market Committee (FOMC) of Federal Reserve.

2.3 Text mining for prediction

On the other hand, more data driven approach could be taken to see if text data has any predictive power of market outcomes: [16] describes such approach as text based market prediction. Most commonly, a group of texts or the so called corpora of texts is gathered, each document is treated as a collection of words (bag-of-words assumption), then features according to some method are extracted from the texts and an outcome (for example categorical outcome of up, down or steady) of the selected financial market/stock is predicted. While it bears some resemblance to sentiment analysis, the way it treats texts and extracts the features is more similar to trying use textual rather than investor sentiment to see if that could predict market outcome. However, the results are mixed: it seems to help predict the volatility of the markets better than the returns but it is still problematic to be able to predict any of the market variable using this approach with high accuracy.

There were also interesting attempts to combine text data analysis methods and orthodox econometric methodology to improve financial forecasting. [25] and [26] employ the so called TEI@I methodology to oil and housing price forecasting with promising results. The methodology combines text mining with econometric time-series forecasting methods to make a combined forecast of a particular market. Unlike previously mentioned approaches, TEI@I puts some structure on the text data in the form of Knowl-

edge Base. That is, the phenomenon of interest is investigated in a top-down fashion: factors that are important are determined and their historic effects are recorded. Then text data is mined using automated algorithms and factors currently at play are investigated. This is combined with econometric trend forecast and a final forecast is generated. So for example if we are modeling price of the oil market, we realize that a war or OPEC embargo historically had X effect on the market price of oil (where X could be simply a range of price movement). Then we use text data to determine which kind of scenario currently is signaled in the retrieved data and then we combine this with ARIMA forecast of the market price. In essence, the idea is to use traditional time-series econometrics for trend forecasting with the abundance of text data for irregular/error component. On the other hand, this requires a thorough investigation and at least some a priori knowledge of the market of interest and restricts the possible information that could be extracted from the text data in a form of a predetermined pattern/scenario.

2.4 Topic modeling in finance

Topic models have seen application in financial forecasting, however most papers focuses on algorithm development/extension rather than analyzing merits of using inferred topics, which are understood to be probability distributions of words (see below). For example, [21] applied Supervised Latent Dirichlet Allocation (LDA) with a goal to predict volatility class of financial firms from their financial reports. While promising, the model did not perform better than the baseline multinomial classification model. More importantly, the paper did not try to investigate further whether estimated topics themselves carry any additional information about the risk of companies described in financial reports.

On the other hand, [6] does exactly that: paper used LDA model in a natural experiment context to see whether there was a change in what Federal Open Market Committee (FOMC) of Federal Reserve was discussing in policy meetings after a requirement to publish transcripts of those meetings has been introduced. The paper shows that LDA is able to estimate topics from FOMC transcripts and that they do seem to be closely related to semantic topics a person would expect to find while reading the transcripts and they do seem to signal a systemic shift in FOMC discussions. One peculiar thing of the paper is though it selects the number of topics in the LDA model in a ad-hoc fashion: K parameter is not data driven but rather selected intuitively.

3 Mathematics of Topic models

So far paper have not yet described Topic models analytically. This section overviews and explicitly describes those topic models which will be used later.

3.1 Latent Dirichlet Allocation

In essence, topic models are probabilistic Bayesian models that by analyzing frequencies of words in text documents are able to represent text documents as mixtures of topics that these texts are about [1]. In short, these models try to model latent semantic structure of text documents. This is especially convenient in financial modeling: it has been acknowledged that media do not only report objective market situation, but by doing so also actively shapes the decisions of market participants [16]. It is reasonable to assume that some news have bigger impact than others. Likewise, topics that news write about should also have different effect.

To get a better sense of how it achieves that, it is reasonable to start with one of the most popular models called Latent Dirichlet Allocation. Informally, we could describe the model as follows [2]: let K be the number of topics that exists in a given corpora, W - number of words in a vocabulary (unique words in a corpora of texts analyzed), α - positive vector of size K and η - a scalar. Also, denote $Dirichlet_V(\alpha)$ V-dimensional Dirichlet distribution with vector α, then:

(1) For each topic, draw a distribution over words: $\beta_k \sim Dirichlet_v(\eta)$
(2) For each document j in D documents:
  i) draw a vector of topic proportions: $\theta_j \sim Dirichlet(\alpha)$
  ii) for each word W:
    a) draw a topic assignment: $z_{jn} \sim Categorical(\theta_j)$
    b) draw a word: $w_{jn} \sim Categorical(\beta_{z_{jn}})$
where $z_{jn} \in \{1, ..., K\}$ and $w_{jn} \in \{1, ..., W\}$.

From this presentation, it is already clear that there are some underlying assumptions behind the model. For example, topics here are just the distributions of words, rather than semantic constructs and they are assumed to be independent of one another which is a strong assumption in this context. As a minor point, number of topics K is assumed to be known in advance which is also an inconvenient feature of the model. On the other hand, the model is able to model documents as consisting of mixture of topics (as com-

pared to having only one topic and being completely unrealistic) and thus despite previous simplifications, have demonstrated interesting applications in various fields [1].

However, the LDA could only best be described as dimensionality reduction model/technique which cannot readily be used for financial market modeling directly because it would be necessary to have some kind of direct connection with the observed outcomes of a particular market. On the other hand, [3] have extended the model to include an additional response variable that depends on the topics in the document θ_j. So now in the generative model we also have another step:

(2) iii) For each document j draw variable y which depends on θ_j

This dependency can be formulated with a Generalized Linear Model or simply through a soft-max (multivariate logistic) function [3]. Now Supervised LDA topic model (SLDA) could be thought of a either classification or a regression model, that takes texts as its input data and produces fitted values for the response variable at the same time inferring the latent topic structure in the corpora.

3.2 Hierarchical Dirichlet Process

One inconvenient feature of LDA is that K or the number of topics is assumed to be known. As a result, a nonparametric counterpart of LDA has been developed. Nonparametric element here is that K is inferred automatically by the model which is called Hierarchical Dirichlet Process mixture model [23] (HDP).

As the name suggests, the model assumes some kind of hierarchical structure and is based on Dirichlet Process. Dirichlet Process itself could be thought of as a measure of measures or a distribution of distributions consisting of discrete elements or atoms as it is customary to call them (for concise exposition of Dirichlet Processes and Hierarchical Dirichlet Processes refer to [23]). The model assumes that there are two Dirichlet Processes, governing global and document level draws of atoms/topic distributions:

$$G_0 \sim DP(\gamma,\ H)$$
$$G_j \sim DP(\alpha_0,\ G_0)$$

where $H \sim Dirichlet(\eta)$ is base (symmetric) distribution $\gamma,\ \alpha_0$ are the

concentration parameters and $j \in \{1, ..., D\}$, where D is the number of documents in the corpus.

Then we draw topics ψ_{jn} associated to n-th word in a document and generate words w_{jn} in the following manner:

$$\psi_{jn} \sim G_j$$
$$w_{jn} \sim Categorical(\psi_{jn})$$

where $w \in \{1, ..., W\}$, and W is the number of words in the vocabulary.

Similarly as in LDA case, HDP model in itself does not have a connection to a response variable. Hence, [27] extends HDP into its supervised counterpart (SHDP). As in SLDA, the response variable is assumed to be drawn after the document topic distributions are drawn and thus:

$$p(y_j|\psi_j, \mu) = f(\psi_j, \mu)$$

where f is assumed to be soft-max (multivariate logistic) function with parameter vector μ (it can be generalized to be Generalized Linear Model).

3.3 Inference of (S)LDA and (S)HDP models

In principle, to estimate the (S)LDA and (S)HDP models, posterior probability of latent variables conditionally on observed data and hyperparameters has to be calculated. For example in LDA case:

$$p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}|\mathbf{w}, \alpha, \eta) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \boldsymbol{\beta}|\alpha, \eta)}{p(\mathbf{w}|\alpha, \eta)}$$

That is, we want to estimate posterior probability of topic distributions $\boldsymbol{\beta}$, document mixture proportions $\boldsymbol{\theta}$ and word-topic assignments $\boldsymbol{z}$ given the observed words in documents $\boldsymbol{w}$ and hyperparameters α and η. Unfortunately, this probability is intractable to compute exactly [2].

To circumvent this problem commonly either some MCMC algorithm or variational inference is used. Since MCMC is can be slow to converge and scales poorly to increase in the size of data, in this paper variational inference algorithms will be used.

Since HDP is basically a generalization of LDA, below some specific HDP variational inference algorithms are presented. Also, HDP inference is much more difficult than LDA inference and it has some specific issues which hinder on its ability to be easily applied (for LDA and SLDA inference refer to [2], [3])

The first peculiar thing with the HDP model is that the Dirichlet Process have several possible representations [23]. Below the so called Chinese Restaurant Franchise metaphor will be described. Thus both for HDP and SHDP that will be used in this paper ([24], [27]), we construct top level Dirichlet process using Sethuraman's Stick-breaking Construction [24]:

$$\begin{aligned}
&\phi_k \sim H \\
&\beta_k^{'} \sim Beta(1, \gamma) \\
&\beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l) \\
&G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}
\end{aligned}$$

where ϕ_k can be interpreted as global topic distributions over vocabulary simplex (probabilities for specific words occurring conditionally on specific topic) and β_k - atom weights for the global Dirichlet process.

Analogously, we apply same construction for the document level process:

$$\begin{aligned}
&\psi_{jt} \sim G_0 \\
&\pi'_{jt} \sim Beta(1, \alpha_0) \\
&\pi_{jt} = \pi'_{jt} \prod_{l=1}^{t-1}(1 - \pi'_{jl}) \\
&G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}
\end{aligned}$$

where similarly to the top level process, ψ_{jt} are document level process atoms (topics) and π_{jt} are atom weights.

To generate the words, two set of indicator variables are drawn:

$$\begin{aligned}
&c_{jt} \sim Categorical(\boldsymbol{\beta}) \\
&z_{jn} \sim Categorical(\boldsymbol{\pi}_j)
\end{aligned}$$

where $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ and $\boldsymbol{\pi}_j = (\pi_{jt})_{t=1}^{\infty}$. These indicators are used in the following fashion:

$$\theta_{jn} = \psi_{jz_{jn}} = \phi_{c_{jz_{jn}}}$$

so that now the words can be generated as $w_{jn} \sim Categorical(\theta_{jn})$

The goal is to estimate the posterior probability posterior probability of latent variables with respect to observed words and hyperparameters.

In a variational inference setting, this intractable posterior (marginal log likelihood of the observed data) is approximated by a variational distribution which is made as close to the posterior as possible by minimizing Kullback-Leibler divergence (for a primer on variational methods, refer to [10]). Thus:

$$\begin{aligned}
\log p(\boldsymbol{w}|\gamma, \alpha_0, \eta) &= \log \int_{\boldsymbol{\beta},\boldsymbol{\pi},\boldsymbol{\phi}} \sum_{\boldsymbol{c},\boldsymbol{z}} p(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\gamma, \alpha_0, \eta) \\
&= \log \int_{\boldsymbol{\beta},\boldsymbol{\pi},\boldsymbol{\phi}} \sum_{\boldsymbol{c},\boldsymbol{z}} p(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\gamma, \alpha_0, \eta) \frac{q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{z})}{q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{z})} \\
&\geq \mathbb{E}_q[\log p(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{z}, \boldsymbol{w}|\gamma, \alpha_0, \eta)] - \mathbb{E}_q[\log q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{z})]
\end{aligned}$$

The q here represents the variational distribution which is assumed to be fully factorized which appeals to mean field variational inference :

$$q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{z}) = q(\boldsymbol{\beta})q(\boldsymbol{\pi})q(\boldsymbol{\phi})q(\boldsymbol{c})q(\boldsymbol{z})$$

The idea is that instead of computing the posterior exactly, somewhat similar distribution - variational distribution is chosen which has much simpler form (breaks down the coupling relationships between latent variables) and is made as close as possible to the true posterior by minimizing the KL divergence.

To minimize this distance, derivatives of the likelihood with respect to each variable is computed and equated to zero which allows to derive variational update equations for each of the variable (for each of the mentioned models, these equations could be found in [2], [3], [24], [27])

Now the main problem with the inference algorithm for HDP model is that while it is postulated in the model that both global and document level Dirichlet Processes have infinite number of atoms, this cannot be implemented directly practically. [24] mitigated this issue by imposing some truncation level: they set K and T to be the maximum number of components (or topics) global and documents processes are allowed to contain. As

such, algorithm is then expected to infer some (possibly) smaller number of topics automatically. Hence, while this truncation requirement is not the same as specifying the number of topics in LDA model, it nevertheless is not efficient.

However, Hughes with different co-authors in a series of papers [4], [8], [7] developed truly automatic algorithm for the HDP model which is able to infer the number of topics from the data in a much more efficient way and is less likely to get stuck in a local minimum. They use the so called Direct Representation and introduces Merge and Delete moves within the inference algorithm. In their approach, instead of specifying truncation levels K and T, the user has to specify only one truncation level K and the Merge and Delete moves removes junk topics so that only the true inferred number of topics are left. As compared to [24] or [27], the final output does not contain junk topics which have to be manually removed by the user. It also makes the inference more efficient because the computation do not use the junk topics. Finally, their approach could use Memoized Online inference, which does not require the specification of a learning rate and can track the global objective function better which means that their approach is a better alternative when huge datasets are considered. For the rest of the paper this model will be referred to as HDPmd.

Appendix A contains derivation of inference algorithm using Direct Representation scheme for the SHDP model which should outperform representation scheme used in [27]. However, in order to avoid the so called research bias, it will not be considered in the further analysis. More specifically, the research started with the goal to develop a well performing prediction mechanism based on the content of texts suited for the financial market prediction. As the research progressed it became clear that in fact the content of text (news in this case) might not be the driving factor and thus topic modeling may not be the way to go if the goal is financial market prediction. On the other hand, topic modeling did allow to infer interesting and quite surprising patterns (see results in next sections) thus the aim of the paper changed and algorithmic development was put into the appendix.

4 Data and empirical approach

4.1 Data description

For the purposes of this paper it is important to find a data source which would nr representative. That is, since the relationship between financial markets and the news is to be investigated, the source of news/news texts

have to be such that the majority of market participants would read them or at least so that it would be relevant for them.

It is customary for well established financial institutions such as investment banks or pension funds to build trading desks which also have inbuilt the so called terminals [18]. The terminals basically allows traders to access the most relevant and important financial information at the real time which includes pricing data, economic data and news articles.

The main dataset in the paper uses exactly this text data from one of the biggest terminals in the market - Thomson Reuters. Basically, the dataset contains every article released in the Thomson Reuters in the period 2008-01-01 to 2009-02-28 (more than 1.8 million).

To complement the articles, external market data is extracted: for the same period values of S&P 500 and VIX (CBOE Volatility Index) indexes are gathered. The choice of these two indexes is simple and intuitive: the first index provides provides a very good representation of US stock market because it consists of 500 largest companies traded on New York Stock Exchange and these companies are traded very frequently (hence, the index itself by market capitalization makes up large proportion of the whole market). The second index is the so called 'fear index' because it measures the implied volatility of S&P 500 index options. Hence, both indexes are related, but the first better depicts the returns of the stock market while the second - volatility of the market. In the text these indexes will be referred to as SP and VIX.

For some experiment also volume and not only index returns of SP will be used.

4.2 Using topic models

Topic models allow to tie these two data sources. For example, at time t (a specific day in the period), we have the released news articles in the Thomson Reuters terminal and from the other data set - the response variable for the market. Hence, we can investigate whether the content of the news articles are predictive or causal (in Granger sense) of financial markets. This is achieved through the estimated topics: after we run some topic model on the dataset, we can infer the topic proportions within the document and thus we basically have K time-series (where K is the estimated number of topics) for our period as well as a time-series for some market variable.

From a data mining perspective, using supervised topic models in this context is easy: we can create pseudo daily documents (treat all documents for a specific day/period t to be one document) and associate each docu-

ment with a response variable (S&P 500 or VIX index value). Then running supervised topic models on such pseudo documents allows directly predict response variable. However, it is important to note that since words in a document is assumed to be exchangeable, synthetic construction of documents could distort the structure of inferred topics and be different to the topics inferred by analyzing real documents.

Using unsupervised topic models in this context is more involved as estimated topics do not provide a straightforward method to predict a market variable. On the other hand, since one part of output from topic models is topic proportions within the documents, these could be used as features to run some classification model and predict the market variables of interest. That is, if we run a topic model on some corpora with K topics, we get a $D \times K$ size matrix, where D is number of documents in the corpora. If documents are aggregated into daily documents, then the number of documents $d \in D$ correspond to numbers of time periods $t \in T$. If not, then normally $D > T$. However, for all documents d released in time period t, we can estimate average topic proportions, so that we force $D = T$ and the same approach could be applied.

Following [16] and [21], we considered Neural Network, Generalized Linear Model with LASSO regularization or Support Vector Machine for presumably best results in this context. However, in practice we found that considering training time and accuracy, Neural Networks and Support Vector Machines were not as good performing as relatively simpler approach using regularized Generalized Linear Model: even though all three approaches were tuned using cross validation, both training time and accuracy in the cases of Neural Network and Support Vector Machine were inferior when compared to regularized General Linear Model. Thus, the former will be used.

5 Ad hoc experiments with data

This section will carry out some ad hoc experiments with the data for the sake of generating some baseline estimates of what we can expect to extract from this dataset.

5.1 Text Data processing

While the Reuters corpus contains huge amount of articles, it is important to try to extract relevant information from noise. For example, the market variables that has been selected are variables for US markets, while the

corpus contains texts not only about US stock markets or stock markets but also sports or geopolitical event around the world. It also contains repeating texts and some technical information bulletins about the terminal itself or tables about some specific market operations which are not suitable for a topic model as an input.

5.1.1 Text document selection

As a result, for this section only those texts that somewhat correspond to US stock markets has been filtered. This was achieved by searching for 'US Stocks' or 'Stock News US' in the title. These keywords correspond to category of articles which are about US stock markets and are quite verbose rather than consisting of tables or other technical information. In other words, these texts are what a human would perceive as a coherent news article text and are suitable for topic modeling.

Also, two options concerning mapping of documents and market variables are considered:

1) selected documents are either aggregated into synthetic daily documents as mentioned previously (all filtered documents in single day are combined into one document) and market response variable is associated directly based on the date.

2) selected documents are not aggregated, but rather, market response variable is associated synthetically, that is for a specific time t which corresponds to some specific day, same market response variable value is assigned to all documents that correspond to time t. Since predictions for market variable will be generated, the most common value predicted for the documents of time t will be selected for market variable in time t (this is possible because the response variables will take discrete value, see below)

5.1.2 Vocabulary selection

Few ad hoc strategies for vocabulary construction will be followed:

1) First approach select only those words that occur in at least 50% of time periods t.

2) Second approach selects all words from the filtered documents

3) Finally, words are ordered by appearance frequency and words that do not correspond to the first and the last quantile are selected. The idea is to discard the most frequent and infrequent words so that these (in a sense) outliers would not affect the inference of topic structure in the corpora.

5.1.3 Vocabulary processing

Also, some transformations of the selected vocabulary will be considered such as stemming and lemmatization. Stemming is the crude rule of thumb that removes the end of the word while lemmatization transforms the variable into the base (dictionary) form [5]. These methods (especially lemmatization) should help to infer topic structure that is similar to true semantic topic structure that a human would perceive as reasonable. For this purpose, Porter's stemmer (SnowballC package for R) and TreeTagger (koRpus package for R) tool for lemmatization is used.

5.2 Market data

As mentioned in previous section, for the external market data, S&P 500 and VIX indexes will be used. However, the values of the indexes are transformed into returns (log-differences) and then these values are discretized - either in two or three categories. For the binary case, the threshold is easy to determine - it is 0, which basically corresponds to the sign of the market variable. In the case where the returns are classified into three categories, a threshold is determine to make three categories approximately equal. While somewhat arbitrarily, this discretization divides market outcomes into upward, downward and steady movement. On the other hand, this arbitrariness does affect results in a major way, because the models are trained with a sample of documents and hence should be able to model the influence of documents to markets in a defined way. Table 1 summarizes the market data:

Table 1: Number of values for market data in each category

Index	Classes	0	1	2
SP	3	104	78	110
SP	2	145	147	
VIX	3	92	97	103
VIX	2	137	155	

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into while '0', '1' and '2' shows how many values each class have.

5.3 Use of topic models

5.3.1 Supervised topic models

For these experiments Supervised Latent Dirichlet Allocation [3] and Supervised Hierarchical Dirichlet Process mixture [27] model will be used with Variational Inference algorithm. For the case when documents are aggregated, the prediction is straightforward. For the case when documents are not aggregated, as mentioned previously, the value of market index at time t is assigned to all documents for time t. Then, supervised topic models are run and prediction generated. Since for time t there will be (possibly) more than one document and thus more than one value is predicted (one for each document), the prediction for time t naturally is selected the one which occurs most often. To run these models, C++ implementation by [27] is used.

5.3.2 Unsupervised topic models and classification models

Unsupervised topic models are somewhat more difficult to apply in this context as their output is only the latent structure of the dataset. Hence, after running these models, we estimate the topic mixtures for each document, that is a distribution of topics within a document and use these mixtures/distributions as a feature set to run some classification algorithm. It closely mimics the idea behind supervised topic models which infer topic structure taking into account the response variable. Interestingly, one possible interpretation of using topic mixtures as a feature set for classification algorithm instead of inferring topic structure conditional on a response variable is that we try to capture the semantic structure of the corpus more closely. That is, inferred topic structure in a supervised topic model setting could possibly be perceived as having less semantic cohesion by a human being because it is guided by a response variable which forces it to infer topics which would be the most predictive of a response variable.

As for the unsupervised topic models, in this section three of them will be used: Latent Dirichlet Allocation [2], Hierarchical Dirichlet Process mixture model [24] and Hierarchical Dirichlet Process mixture model with Merge and Delete moves [8], [7]. Basically, all models uses Variational Inference algorithms, but the last one offers truly automatic inference of number of topics K and uses different assignment scheme compared to [24]. These models are implemented in Gensim or BNPY packages available for Python.

For the classification algorithms, following [21] and [16] regularized Generalized Linear Model (GLM) with LASSO penalty is used. The model are

tuned using 10-fold cross validation over some grid of parameters. It is implemented in glmnet package for R.

5.4 Hyperparameters of topic models

It is important to stress one particular feature of this ad hoc experiment - basically, there will be very little attention paid to tuning the hyperparameters of the topic models. Only attention will be paid to the number of topic paramter K and word-topic distribution concentration parameter α. Only HDP model with inference scheme by [7], [8] can infer K automatically, so first of all, this model is run with Merge and Delete moves for which we specify some upper bound for K (normally - 200) and let the model infer smaller number of active atoms. Then we use this number for SLDA, SHDP, LDA and HDP models. In SHDP and HDP models, we set both global level topic truncation and document level topic truncation parameters to this value. For the α, we set 0.1 for these experiments and leave the rest to be set by default values of the packages that we use.

5.5 Summary of combinations and approach

So the main idea is to take a processed text data dataset, combine it with some market index, run a supervised or unsupervised topic model and then generate predictions for the unseen documents. That is, 292 days are available for which there are some text documents . 30 of the last days or time periods are left for prediction, so the topic models are run for documents covering first 262 days. This amounts to about 10% of the dataset. This choice was governed by the fact that this approach is more similar to data mining or data driven analysis as compared to classical statistical analysis. That is, in sample analysis is not a reasonable metric for capturing the predictive relationship between text data and market data. On the other hand, while the topic model will not use all of the text documents for model training, this is not the same as forecasting in a regular time-series analysis, because these unseen documents will be fed into a trained model to estimate topic mixtures within the documents and only then the market variable predictions will be generated. So this whole approach could be best interpreted as analysing whether the content of news are descriptive of the events and whether it actually corresponds to outcomes of a specified market variables.

Summarizing the previous subsections:

- Specific documents will be filtered: those related to US stock markets will be selected

- Documents will be either aggregated into pseudo-daily documents or not
- For vocabulary construction, words will be selected appearing in at least 50% of time periods t (will be denoted by A), all words from selected documents (denoted by N) or those within second and third quantile most frequently occurring)denoted by Q)
- Selected vocabulary will be either unprocessed or lemmatized or both lemmatized and then stemmed.
- For market data, S&P 500 (SP) and VIX indexes will be used and values of returns of these indexes will be discretized into 2 or 3 classes
- For classification models, either SLDA, SHDP models will be used, or LDA, HDP, HDP with Merge and Delete (HDPmd) moves along with regularized GLM.
- All models will be run for 10 times and accuracy from the validation set of unseen document averaged.

5.6 Results

First of all, it is important to note that SHDP model (or at least implementation by [27]) did extremely poorly in terms of its applicability: the computations were run on few servers for couple of weeks and even though the number of documents was not particularly large, for some of the combinations SHDP was not able to finish to run all of 10 cycles (ar even at least one). Thus, computation were stopped and we report the results for SHDP for those combinations for which the model was able to finish.

Tables 2 and 3 report the results averaged over all the previously specified combinations. From these tables the first thing to note is how small the difference between the accuracy of unsupervised topic models with GLM for market condition classification is compared to supervised topic models. Also, it is already evident, how poorly HDP (which does not use Direct Representation scheme) model performs in both - supervised and unsupervised setting.

Table 2: Average accuracy of predicting market variables using unsupervised topic models with GLM

Index	Classes	Model	Random	Accuracy	CV	N
SP	3	HDP	0.233	0.351	0.312	180
SP	3	HDPmd	0.233	0.494	0.253	180
SP	3	LDA	0.233	0.469	0.267	180
SP	2	HDP	0.567	0.485	0.108	180
SP	2	HDPmd	0.567	0.577	0.233	180
SP	2	LDA	0.567	0.615	0.204	180
VIX	3	HDP	0.333	0.325	0.159	180
VIX	3	HDPmd	0.333	0.438	0.320	180
VIX	3	LDA	0.333	0.444	0.280	180
VIX	2	HDP	0.533	0.496	0.095	180
VIX	2	HDPmd	0.533	0.611	0.234	180
VIX	2	LDA	0.533	0.634	0.221	180

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into; 'Random' is the the accuracy of selecting the most frequent value in the training set for prediction; 'CV' - Coefficient of Variation and 'N' is the number of observations for averaging.

Table 3: Average accuracy of predicting market variables using supervised topic models

Index	Classes	Model	Accuracy	CV	N
SP	3	SHDP	0.399	0.228	51
SP	3	SLDA	0.500	0.225	180
SP	2	SHDP	0.526	0.146	50
SP	2	SLDA	0.626	0.182	180
VIX	3	SHDP	0.374	0.213	51
VIX	3	SLDA	0.456	0.260	180
VIX	2	SHDP	0.548	0.170	49
VIX	2	SLDA	0.641	0.205	180

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into; 'CV' - Coefficient of Variation and 'N' is the number of observations for averaging.

Since we have quite a few combinations of how to preprocess data, it is important to see how the accuracy differed among differently prepared datasets, thus tables 4 and 5 shows averaged accuracy (along with coefficients of variation) over all the run models and market variables.

Table 4: Average accuracy of predicting market variables for differently processed subsets of data using unsupervised topic models

Agg	Lemm	Stemm	Voc	Random	Accuracy	CV	N
Yes	Yes	Yes	A	0.417	0.612	0.322	120
Yes	Yes	No	A	0.417	0.602	0.333	120
Yes	No	No	A	0.417	0.584	0.304	120
Yes	Yes	Yes	N	0.417	0.557	0.301	120
Yes	Yes	No	N	0.417	0.540	0.315	120
No	Yes	Yes	A	0.417	0.531	0.280	120
No	Yes	No	A	0.417	0.518	0.252	120
No	Yes	Yes	N	0.417	0.508	0.238	120
No	Yes	No	N	0.417	0.492	0.249	120
No	No	No	N	0.417	0.468	0.217	120
Yes	No	No	N	0.417	0.462	0.261	120
No	No	No	A	0.417	0.458	0.207	120
No	No	No	Q	0.417	0.452	0.217	120
No	Yes	No	Q	0.417	0.450	0.240	120
No	Yes	Yes	Q	0.417	0.442	0.266	120
Yes	Yes	Yes	Q	0.417	0.414	0.235	120
Yes	Yes	No	Q	0.417	0.413	0.243	120
Yes	No	No	Q	0.417	0.405	0.241	120

Notes: Columns 'Agg', 'Lem', 'Stem', 'Voc' denote respectively whether the documents were aggregated or not, whether the vocabulary was lemmatized or nor, whether the vocabulary was stemmed or not and the type of method for selecting words (see subsection 5.5); 'Random' - accuracy of prediction when the most frequent value in the training set is selected; 'CV' - Coefficient of Variation and 'N' - number of observations for averaging. The results are ordered by the 'Accuracy' column in a decreasing sort.

Note: due to previously mentioned fact that SHDP model did not run for some of the combinations, in this summary table supervised topic models seem to perform much better. However, it just mostly reflects the fact that for the best performing combinations, only SLDA model managed to run. On the other hand, both table points to the same type of methods for processing data. That is, apparently in both cases, the accuracy is the highest when we aggregate documents into daily documents and use words that occurs in at least 50 time periods. Also, lemmatization and lemmatization with stemming helps. On the other hand, while average accuracy is considerably higher compared to other methods, it raises a question, whether

the content of news correspond to market outcomes or rather some specific, often occurring words tend to coincide with market conditions.

Table 5: Average accuracy of predicting market variables for differently processed subsets of data using supervised topic models

Agg	Lem	Stem	Voc	Accuracy	CV	N
Yes	Yes	Yes	A	0.753	0.099	40
Yes	Yes	No	A	0.705	0.145	40
Yes	Yes	Yes	N	0.625	0.223	40
No	Yes	Yes	A	0.619	0.151	40
Yes	No	No	A	0.617	0.190	78
No	Yes	No	A	0.610	0.154	40
No	Yes	Yes	N	0.606	0.136	40
No	Yes	No	N	0.600	0.141	40
Yes	Yes	No	N	0.587	0.255	40
No	No	No	N	0.508	0.222	43
No	No	No	A	0.501	0.221	80
Yes	No	No	N	0.477	0.212	80
No	Yes	Yes	Q	0.458	0.214	40
No	No	No	Q	0.455	0.234	80
No	Yes	No	Q	0.447	0.193	40
Yes	Yes	No	Q	0.431	0.242	40
Yes	Yes	Yes	Q	0.427	0.222	40
Yes	No	No	Q	0.422	0.309	80

Notes: Columns 'Agg', 'Lem', 'Stem', 'Voc' denote respectively whether the documents were aggregated or not, whether the vocabulary was lemmatized or nor, whether the vocabulary was stemmed or not and the type of method for selecting words (see subsection 5.5); 'CV' - Coefficient of Variation and 'N' - number of observations for averaging. The results are ordered by the 'Accuracy' column in a decreasing sort.

This doubt is further strengthened by another observation: both type of topic models performs the worst when we drops the most frequent and infrequent words. It is quite natural to think that the semantic content of human speech is not determined by the most frequently/infrequently used words, but rather those which differentiate the topics we communicate about. Hence, already at this stage it is questionable, whether the thematic content of news correspond to market outcomes.

On the other hand, to make a more fair comparison between performance of supervised and unsupervised topic modeling approaches, tables 6 and 7

compares accuracy of both approaches for the the dataset that the average over models' accuracy was highest (aggregated documents, stemmed and lemmatized vocabulary, words occuring in at least 50% of time periods).

Table 6: Average accuracy of predicting market variables for best performing data processing method and unsupervised topic models

Index	Classes	Model	Random	Accuracy	CV	N
SP	3	HDP	0.233	0.247	0.223	10
SP	3	HDPmd	0.233	0.727	0.071	10
SP	3	LDA	0.233	0.620	0.058	10
SP	2	HDP	0.567	0.450	0.087	10
SP	2	HDPmd	0.567	0.780	0.030	10
SP	2	LDA	0.567	0.807	0.026	10
VIX	3	HDP	0.333	0.303	0.204	10
VIX	3	HDPmd	0.333	0.690	0.094	10
VIX	3	LDA	0.333	0.620	0.058	10
VIX	2	HDP	0.533	0.477	0.114	10
VIX	2	HDPmd	0.533	0.823	0.043	10
VIX	2	LDA	0.533	0.807	0.033	10

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into; 'Random' is the the accuracy of selecting the most frequent value in the training set for prediction; 'CV' - Coefficient of Variation and 'N' is the number of observations for averaging.

Table 7: Average accuracy of predicting market variables for best performing data processing method and supervised topic models

Index	Classes	Model	Accuracy	CV	N
SP	3	SLDA	0.723	0.053	10
SP	2	SLDA	0.777	0.067	10
VIX	3	SLDA	0.670	0.049	10
VIX	2	SLDA	0.843	0.019	10

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into; 'CV' - Coefficient of Variation and 'N' is the number of observations for averaging.

Now it is evident that using only unsupervised topic models with some classification mechanism seems to perform better compared to supervised topic models in most cases (if we consider LDA or especially HDPmd model with GLM used for classification). In fact, this trend is evident for other processing schemes as well.

We end this section with another interesting opservation. Since it is easy to incorporate other features into GLM used with unsupervised topic models, we regress market variables onto estimated topics and textual sentiment, following [22], [17]: we use the prepared word lists indicating anxiety or excitement and construct relative sentiment index which was mentioned earlier in the text. Table 8 shows how the average accuracy changes when textual sentiment is included. Column 'Sentiment' lists what textual sentiment measure was used: either none, index as in [17] or just two separate variables for excitement and anxiety words' occurrences normalized by the character count of the document.

Table 8: Average accuracy of predicting market variables using HDPmd model additionally including features capturing textual sentiment

Index	Classes	Sentiment	Random	Accuracy	CV	N
SP	3	None	0.233	0.494	0.253	180
SP	3	Index	0.233	0.515	0.242	180
SP	3	Variables	0.233	0.508	0.246	180
SP	2	None	0.567	0.577	0.233	180
SP	2	Index	0.567	0.674	0.167	180
SP	2	Variables	0.567	0.673	0.155	180
VIX	3	None	0.333	0.438	0.320	180
VIX	3	Index	0.333	0.526	0.199	180
VIX	3	Variables	0.333	0.516	0.198	180
VIX	2	None	0.533	0.611	0.234	180
VIX	2	Index	0.533	0.703	0.165	180
VIX	2	Variables	0.533	0.697	0.155	180

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into; 'Sentiment' shows what kind of features was incorporated to capture textual sentiment (see paragraph above);'Random' is the the accuracy of selecting the most frequent value in the training set for prediction; 'CV' - Coefficient of Variation and 'N' is the number of observations for averaging.

Interestingly, textual sentiment index does seem to improve average accuracy quite substantially, though the absolute accuracy is still somewhat low considering the setting of the whole experiment. On the other hand that points to a possiblity, that not only thematic content might be related to market outcomes but the way the news are presented or expressed.

6 Content of news and behavior of markets

In this section we take a more structured approach to analyzing the relationship between the content of news and outcomes of financial markets. Specifically, we use will previous section's results as a guide to selecting type of models to use and as a baseline to what kind of predictive accuracy could be expected. Also, unlike in the previous section, topic models will use previously empirically tested and suggested prior values for the hyperparameters.

The idea in this section is that since the selection of hyperparameters for topic modeling is more careful and thus we have a bigger probability to actually characterize the thematic content of news (rather than arbitrary clusters of words), we are going to investigate not only whether content of news are descriptive of market behavior but also whether it is causal in a Granger sense.

6.1 Text Data Processing

Now the preprocessing of the text data will play a less important role. Also, results of previous section will serve as a guide to shaping experiment design decisions.

As in previously, 3 steps are considered:

- Document selection: two subsets of corpus will be considered - one with texts related to US stocks specifically, as previously; and another one with basically all the texts in the corpus except for duplicates, technical tables and etc. as not selected in a subsetting procedure based manual inspection of the corpus. In the first case, 3193 documents are selected while in the second - 615920 documents.

- Document aggregation: unlike previously, documents will not be synthetically aggregated because now the focus of the experiment is to estimate topic structure which would be as close to the semantic one as possible. Instead, as later topic mixtures within documents will be

used for classification or Granger causality testing, the topic mixture for time t will be estimated by averaging. That is, while many more documents and topic mixtures will be available for each time t, average topic mixtures within these document for each time period will be estimated.

- Vocabulary selection: [17] will be closely followed . Here words are ranked in their ability to dsicriminate among documents: term-frequency inverse-document-frequency is computed and the scores plotted to determine a threshold. For the first case this is equal to 10 while in the second - 21. This approach punishes words when they are infrequent or appears in many documents. Also, a list of stopwords is used. In the first case this results in 3630 words in a vocabulary, while in the second - 36992 words.

- Vocabulary will be lemmatized but not stemmed, which produced similar results in previous section however, lemmatization seems to be more theoretically sound.

Finally, textual sentiment is also extracted from the selected subsets of data using approach in [22] and [6] and either the relative sentiment index is constructed or normalized values for anxiety and excitement is used as separate features for classiciation.

6.2 Market data

As in previous section, S%P500 and VIX (SP, VIX) indexes will be used for response variable. Additionally, a measure of market volume of trading will be incorporated as well. This is captured by using volume of S&P500 index trading. As in previous case, it is discretized using the same logic: log-differences of the volume is dsicretized either into 3-classes or into binary representation. Table 9 shows distribution of values for these variables.

When testing for Granger causality, the market data will not be discretized. Specifically, time-series for S&P500, volume of S&P500 and VIX will be used in its absolute values (for details see below).

6.3 Selection of topic models and hyperparameters

Since the previous section revealed that in this context there is little difference between supervised and unsupervised topic models (while in fact unsupervised perfomed even better in our experiments) for this section we

Table 9: Number of values for volume market data in each category

Index	Class	0	1	2
SP (Volume)	3	98	100	94
SP (Volume)	2	150	142	

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into while '0', '1' and '2' shows how many values each class have.

are going to use two of the best performed unsupervised topic models: LDA and HDP with Merge and Delete moves.

Differently from before, two sets of hyperparameters will be used following [17] and [7]:

- The first set uses $\alpha = 0.5$, $\eta = 0.1$ and $\gamma = 10$
- The second set $\alpha = 50/K$ (where K is the number of topic), $\eta = 0.025$ and $\gamma = 10$

Second set follows [17] which similary to this paper tried to characterize thematic content of some economic/financial corpora, however, used LDA model and thus did not need to specify value for γ. The first set is taken from [7] used HDP model and so had to set all the values but they focused on more generic or even synthetic corpora.

Note, when the LDA model in the experiment is estimate, the γ parameter is ignored because it is not in the model construction. Hence, γ is relevant only for HDPmd model. Also note that like previously, first of all HDPmd model will be estimated first and inferred number of topics K will be averaged and used as a parameter for the LDA model.

6.4 Overview of experiments

As mentioned in the beginning of this section, basically two different kinds of experiments will be executed: one for analyzing descriptive/ predictive relationship between financial markets and financial news and one for analyzing causal (in a Granger sense) relationship between the two.

For the first the analysis follows in the same manner as in previous section. That is, two topic models with two sets of hyperparameters on two subsets of corpora is run and then topic mixtures within documents is inferred. Average topic proportions within documents for time period t is

estimated, so that number of mixtures is equal to time periods (days).This set of features is used with two classification models (trained) on 7 different response variables. Also, these models are trained with sentiment index, without it or incorporating normalized variables for textual excitement and anxiety. Then the accuracy is inspected on the set of unseen documents, just like in the previous section. For the smaller subset of data, 10 cycles will be run while for the larger - 5 cycles as it contains more than 600000 documents and thus computation becomes expensive.

For the second analysis we use infered topic mixtures within documents to estimate Granger causality between estimated topics and market variables to see whether some specific topics/content of financial news Granger cause market outcomes, the reverse is true or neither.

Specifically, since we have some N_t documents for each day t and run selected topic models, we get some K topics (distributions). As in previous experiments, the parameters of topic models also lets us to infer these K topic proportions within each documents. Since N_t is (normally) much larger than 1 and so there are more than one document for each data point of market data, we estimate average topic proportions within each period t. Now denote X_{tk} the time series for a specific topic $k \in K$. Basically, it is the proportion of topic k at time t (time periods cover 262 trading days in the training sample). Also, denote Y_t - the market variable of interest: either S&P500, volume of S&P500 or VIX index. Then we can test for absence of Granger causality using the following VAR model:

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = A_0 + A_1 \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \cdots + A_p \begin{pmatrix} X_{t-p} \\ Y_{t-p} \end{pmatrix} + U_t$$

where A_0 is 2×1 matrix of intercepts, $A_1, \ldots, A_p$ are 2×2 matrices of coefficients and U_t is 2×1 matrix of errors. Now if specific coefficients of matrices $(A_1)_{12}, \ldots, (A_p)_{12}$ are all equal to zero, then it means Y_t does not Granger cause X_t. This hypothesis could be tested statistically. We use the Todo-Yamamoto [9] procedure. As such, we can then tell which Granger causes which - markets or the news (specific topics of the news).

6.5 Results of first experiment

Tables 10 and 11 basically summarizes the classification results for both datasets - using texts related to only US Stocks and all filtered texts for different response variables.

Table 10: Average accuracy of predicting market variables using subset of data related to US stocks/markets

Index	Classes	Model	H. Set	Random	Accuracy	CV	N
SP	3	HDPmd	1	0.233	0.440	0.244	10
SP	3	HDPmd	2	0.233	0.540	0.136	10
SP	3	LDA	1	0.233	0.477	0.148	10
SP	3	LDA	2	0.233	0.467	0.175	10
SP	2	HDPmd	1	0.567	0.600	0.141	10
SP	2	HDPmd	2	0.567	0.657	0.124	10
SP	2	LDA	1	0.567	0.593	0.124	10
SP	2	LDA	2	0.567	0.647	0.162	10
SP (Volume)	3	HDPmd	1	0.567	0.390	0.091	10
SP (Volume)	3	HDPmd	2	0.567	0.403	0.099	10
SP (Volume)	3	LDA	1	0.567	0.413	0.109	10
SP (Volume)	3	LDA	2	0.567	0.373	0.110	10
SP (Volume)	2	HDPmd	1	0.300	0.550	0.170	10
SP (Volume)	2	HDPmd	2	0.300	0.583	0.090	10
SP (Volume)	2	LDA	1	0.300	0.587	0.072	10
SP (Volume)	2	LDA	2	0.300	0.560	0.153	10
VIX	3	HDPmd	1	0.333	0.407	0.133	10
VIX	3	HDPmd	2	0.333	0.500	0.122	10
VIX	3	LDA	1	0.333	0.477	0.206	10
VIX	3	LDA	2	0.333	0.460	0.159	10
VIX	2	HDPmd	1	0.533	0.620	0.130	10
VIX	2	HDPmd	2	0.533	0.660	0.062	10
VIX	2	LDA	1	0.533	0.603	0.118	10
VIX	2	LDA	2	0.533	0.657	0.120	10

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into; 'H. Set' marks which set of hyperparameters was used (see subsection 6.3);'Random' is the the accuracy of selecting the most frequent value in the training set for prediction; 'CV' - Coefficient of Variation and 'N' is the number of observations for averaging.

Table 11: Average accuracy of predicting market variables using all filtered documents

Index	Classes	Model	H.Set	Random	Accuracy	CV	N
SP	3	HDPmd	1	0.233	0.233	0	5
SP	3	HDPmdP	2	0.233	0.233	0	5
SP	3	LDA	1	0.233	0.287	0.303	5
SP	3	LDA	2	0.233	0.393	0.126	5
SP	2	HDPmd	1	0.567	0.493	0.074	5
SP	2	HDPmd	2	0.567	0.467	0	5
SP	2	LDA	1	0.567	0.467	0	5
SP	2	LDA	2	0.567	0.507	0.118	5
SP (Volume)	3	HDPmd	1	0.567	0.333	0.187	5
SP (Volume)	3	HDPmd	2	0.567	0.373	0.160	5
SP (Volume)	3	LDA	1	0.567	0.400	0	5
SP (Volume)	3	LDA	2	0.567	0.387	0.131	5
SP (Volume)	2	HDPmd	1	0.300	0.400	0.059	5
SP (Volume)	2	HDPmd	2	0.300	0.393	0.038	5
SP (Volume)	2	LDA	1	0.300	0.600	0	5
SP (Volume)	2	LDA	2	0.300	0.567	0.102	5
VIX	3	HDPmd	1	0.333	0.327	0.046	5
VIX	3	HDPmd	2	0.333	0.340	0.044	5
VIX	3	LDA	1	0.333	0.333	0	5
VIX	3	LDA	2	0.333	0.320	0.228	5
VIX	2	HDPmd	1	0.533	0.500	0.125	5
VIX	2	HDPmd	2	0.533	0.473	0.031	5
VIX	2	LDA	1	0.533	0.467	0	5
VIX	2	LDA	2	0.533	0.467	0	5

Notes: 'Classes' marks the number of classes 'Index' variable was divided (discretized) into; 'H. Set' marks which set of hyperparameters was used (see subsection 6.3);'Random' is the the accuracy of selecting the most frequent value in the training set for prediction; 'CV' - Coefficient of Variation and 'N' is the number of observations for averaging.

The first thing that is noteable is that compared to the ad-hoc experiments previously, the accuracy of the models compared to the random guess is not better by using the models properly (using suggested hyperparameter values, not synthetically aggregating documents, etc.). That is, if we try to model the content of news and use it to predict the state of the market for

some variables - the results disappoint. This situation is not made better by using the full (filtered) dataset as portrayed in the table 11: in fact, the accuracy seems to be worse for most cases. This is does not seem impossible because huge amount of texts also contain much more noise or useless news regarding the markets of interest.

Another point is that there does not seem to be huge difference between different sets of hyperparameters: in the US stcoks texts case, the second set seems to perform better while when all filtered texts are used, somewhat the first set give a little edge though in overall with this dataset the models barely outperforms random guesses if at all.

Also, while not shown in the text, inclusion of sentiment features had comparable effect to the one in ad hoc experiments: while not substantially but increased the average accuracy if included either as an relative sentiment index or separate features.

While experiments carried out currently do not prove that the content of the new itself are useless for predicting the state of the market as the approach might have some flawed properties, the way these experiments were set up does somewhat support the idea that the content or 'what is written in the news' does not matter for the markets that much. Primarily because the prediction in this paper is really an artificial one. That is, we assume that we know the all the necessary texts published in a representative source (Thomson Reuters terminal) at time t and by using them we try to assign a value for the market variable of the same time period. So we are not forecasting into the future, but rather see how well the variation in topic mixtures in documents correspond to states of market variables. As topic modeling have seen successful application in other contexts where topic structure (resembling semantic topics that humans think about) where had to be modeled and used for example for classification of documents, we see the outcomes of these experiments as pointing to the fact that actually, the content of the news is not that all important for the market participants.

6.6 Results of the second experiment

Table 12 and 13 summarizes the Granger causality testing between market variables and topic mixtures within documents.

Table 12: Granger causality testing between market variables and inferred topic mixtures for the dataset related to US stocks

Model	H. Set	Index	Topic Cause	CV	Index Cause	N
HDPmd	1	SP	16	0.437	0	105
HDPmd	1	SP (Volume)	20	0.527	0	112
HDPmd	1	VIX	7	1.355	0	87
HDPmd	2	SP	16	0.527	0	185
HDPmd	2	SP (Volume)	27	0.782	0	189
HDPmd	2	VIX	16	0.734	0	169
LDA	1	SP	9	0.820	0	98
LDA	1	SP (Volume)	15	0.567	0	105
LDA	1	VIX	15	0.471	0	91
LDA	2	SP	19	0.677	0	174
LDA	2	SP (Volume)	31	0.355	0	188
LDA	2	VIX	9	0.820	0	158

Notes: 'H. Set' marks which set of hyperparameters was used (see subsection 6.3); 'Topic Cause' shows number of instances, when a time-series for specific topic was found to Granger cause 'Index' variable, while 'Index Cause' is the opposite case; 'CV' is the Coefficient of Variation for the 'Topic Cause' variable over the different runs of models and 'N' is the number of tests.

The tables sums the number of instances (over run iterations for different models and datasets) that either topic time-series were Granger causal of some market index variable or either the index variable was Granger causal of some topic mixtures time-series. Since procedure for testing the Granger causality requires a well specified VAR model, cases for which it was impossible to find a specification of VAR that would pass the regular statistical tests (for example: size of roots, absence of auto-correlation in the residuals, etc.) were discarded.

Table 13: Granger causality testing between market variables and inferred topic mixtures for the dataset with all filtered texts

Model	H. Set	Index	Topic Cause	CV	Index Cause	N
HDPmd	1	SP	41	0.468	0	778
HDPmd	1	SP (Volume)	40	0.234	0	786
HDPmd	1	VIX	37	0.181	0	752
HDPmd	2	SP	50	0.324	0	806
HDPmd	2	SP (Volume)	44	0.354	0	811
HDPmd	2	VIX	40	0.177	0	779
LDA	1	SP	74	0.193	0	716
LDA	1	SP (Volume)	155	0.094	0	760
LDA	1	VIX	86	0.171	0	707
LDA	2	SP	63	0.255	0	731
LDA	2	SP (Volume)	135	0.083	0	777
LDA	2	VIX	65	0.218	0	735

Notes: 'H. Set' marks which set of hyperparameters was used (see subsection 6.3); 'Topic Cause' shows number of instances, when a time-series for specific topic was found to Granger cause 'Index' variable, while 'Index Cause' is the opposite case; 'CV' is the Coefficient of Variation for the 'Topic Cause' variable over the different runs of models and 'N' is the number of tests.

What is evident at first sight is that no instances were found for market index variable to Granger cause any of the inferred topics for any of the dataset or model. On the other hand, there we quite a few of instances when inferred specific topic seemed to granger cause market index variable. What seems troubling is the variation in the instances of topic causing a market variable when averaged over different runs of the models. While especially evident in the US Stocks dataset, this rather big variation could easily be explained: this experiment uses topic mixtures within time periods as time-series for testing Granger causality and market variables. As such, this dataset provides little space for irrelevant topics to be inferred because the texts covers US stock news. Furthermore, the dataset are rather limited in size and since topic models uses random values for initialization of inference algorithm, small changes in inferred topic mixtures can lead to bigger variation in the dynamics of the time series. Importantly, used topic models do not model topics in a dynamic fashion, and so these time series are sensitive.

On the other hand, when we use big dataset of texts having coverage on much wider set of topics, inferred topic mixtures are less sensitive to initialization values and somewhat more stable. Hence, table 12 should be taken with a handful of salt.

However, both tables seem to point to the same direction: some topics rather index variable seem to carry the causality in the Granger sense and apparently, the volume of trading seems to be caused by some specific topics the most as compared to returns or volatility.

Of course, in this context a lot of spurious Granger causality could have been detected. Figures 1 to 4 provide evidence, that it is not necessarily the case. It depicts most probable words for inferred topics that seem to Granger cause the volume of S&P 500 index trading using the dataset with all filtered documents over 5 runs of LDA model with second set of hyperparameters.

Figure 1: Topic about Federal Reserve

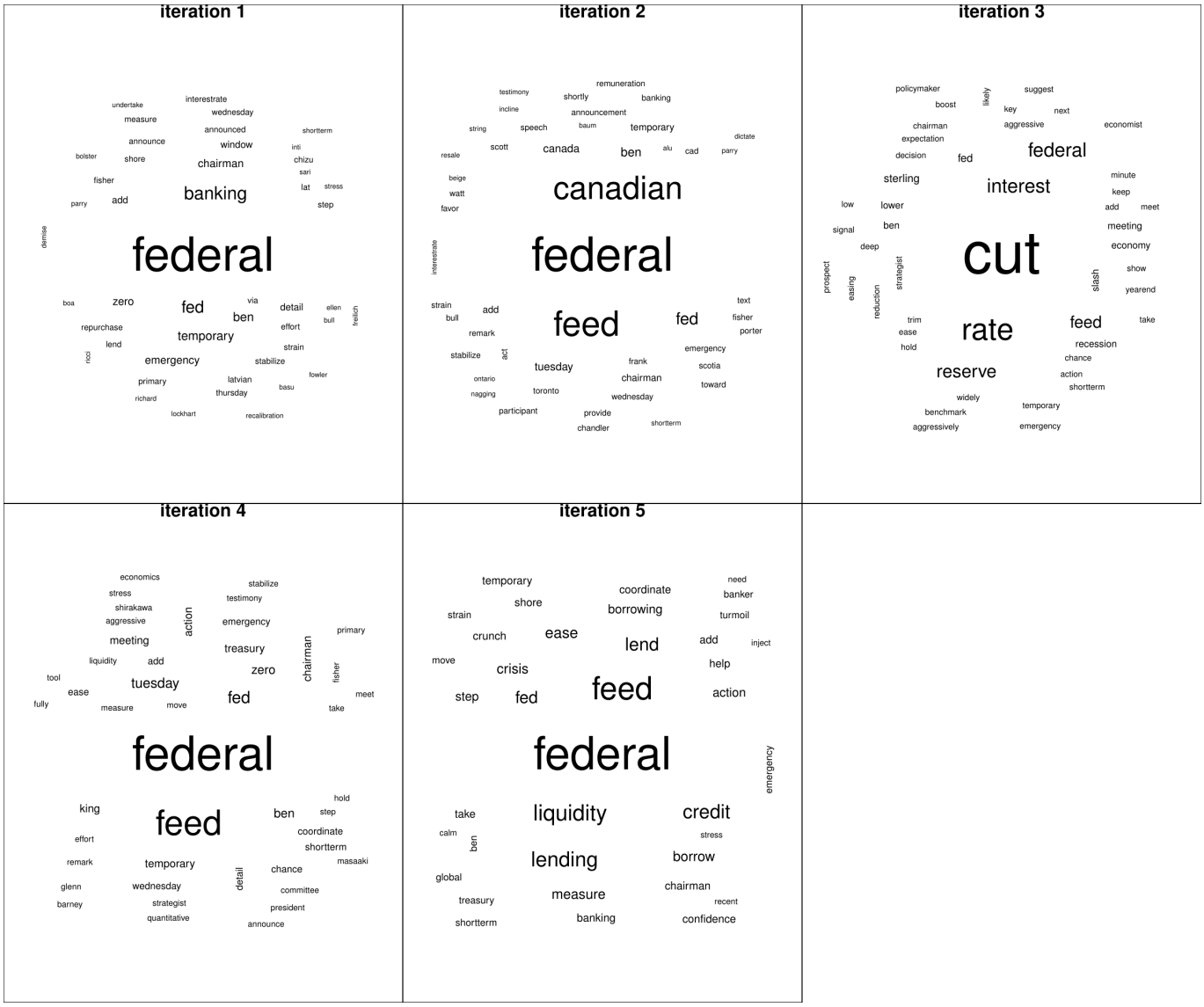


Figure 2: Topic related to rating agencies

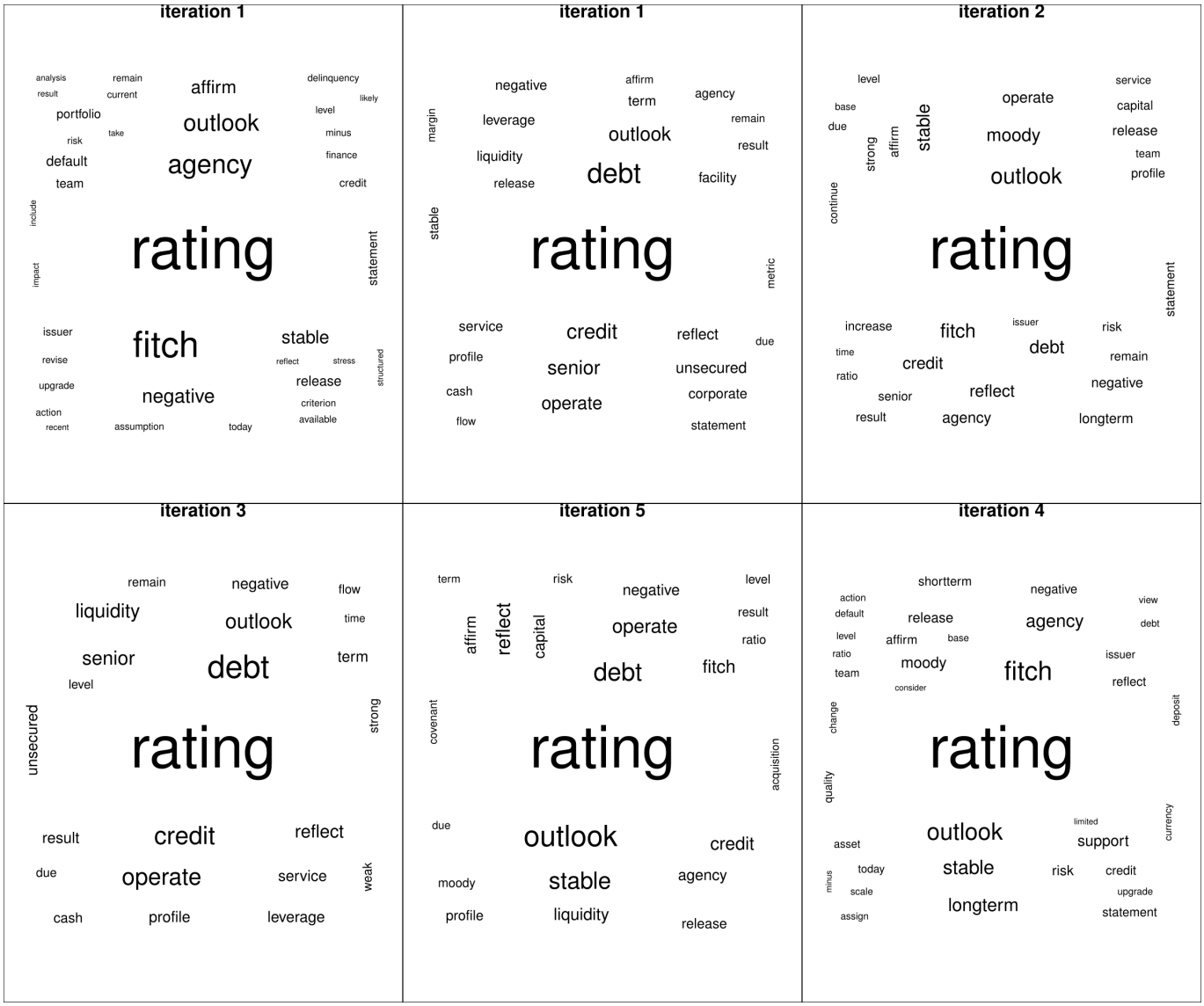


Figure 3: Topic about economic situation/slowdown

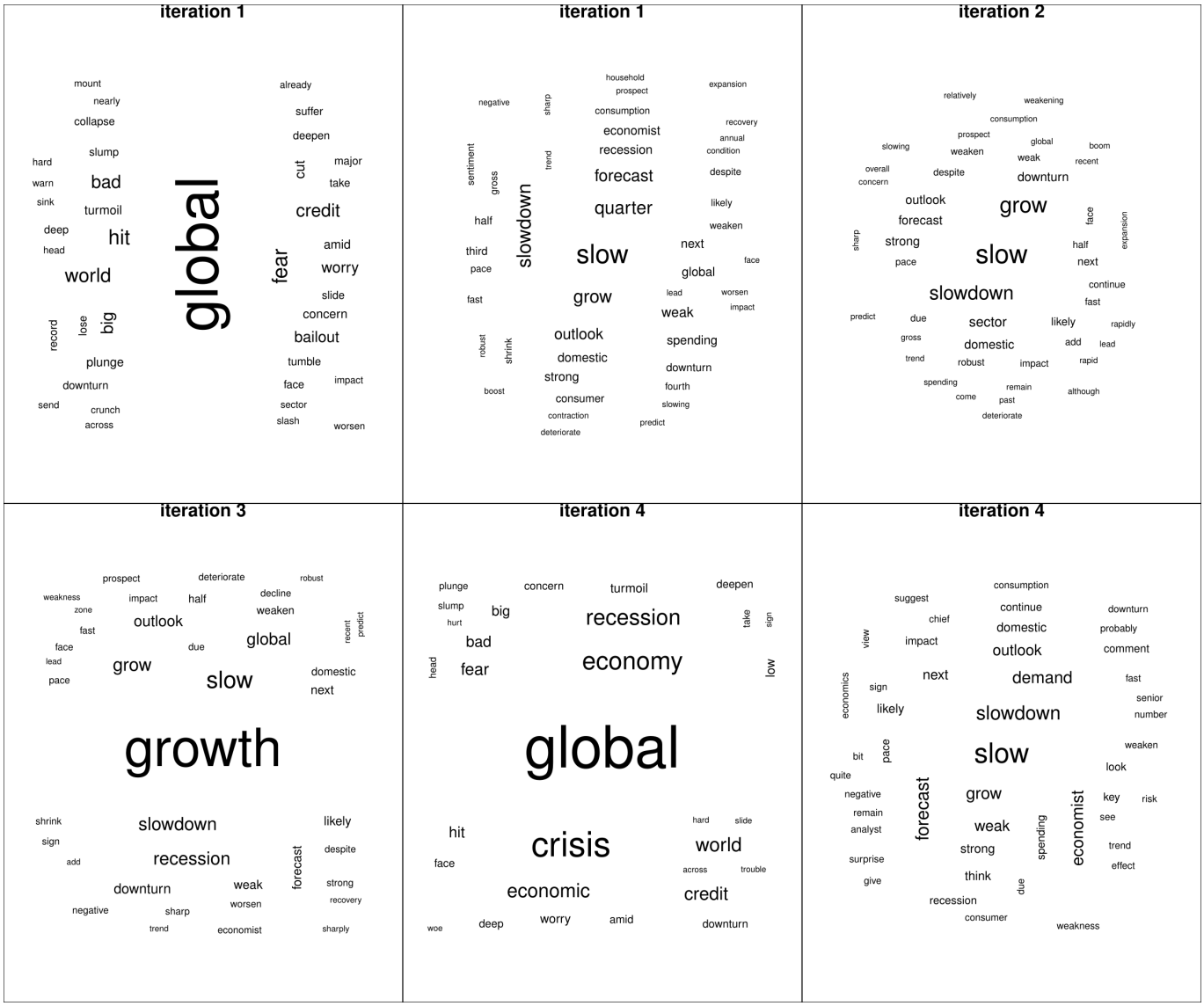
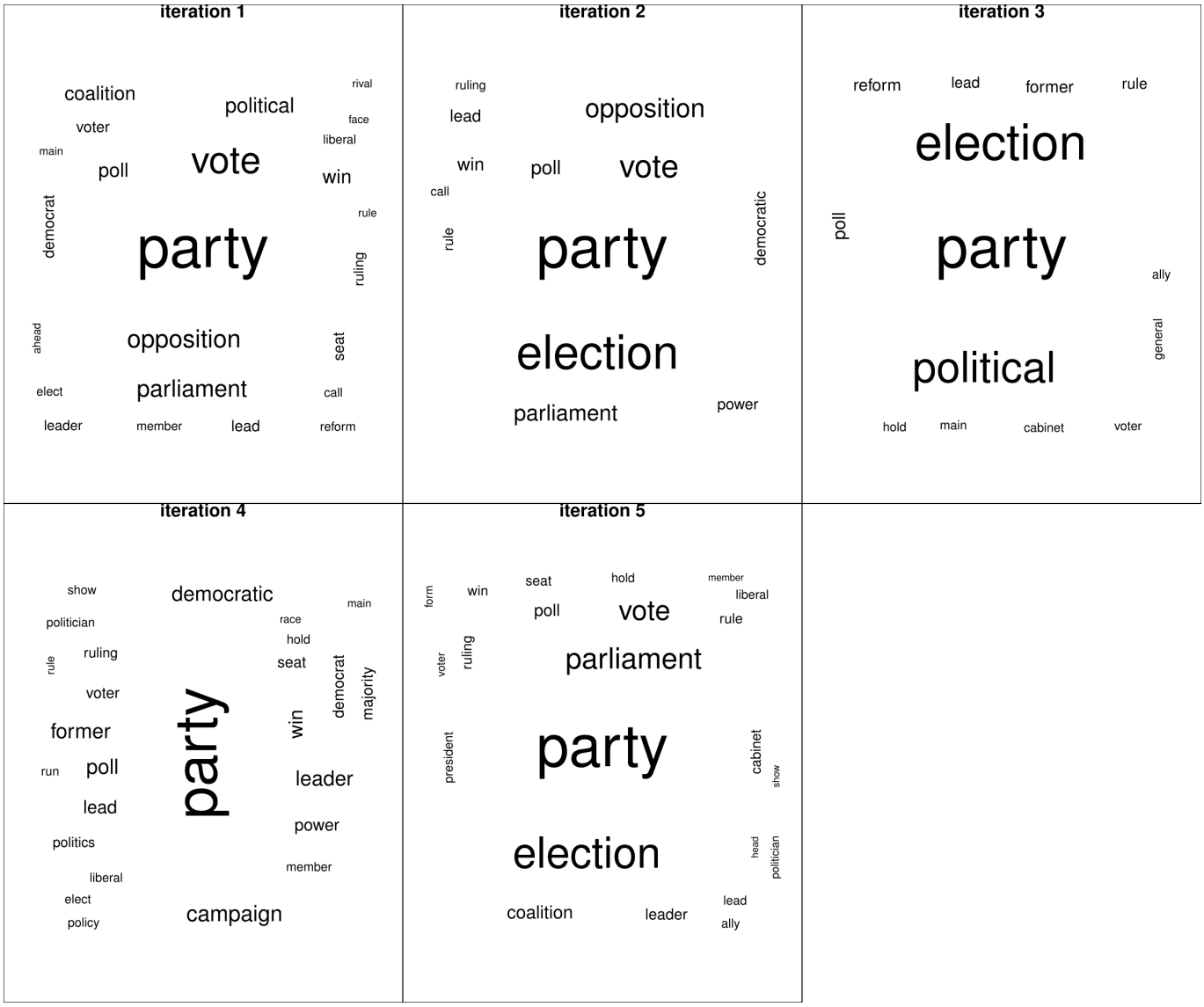


Figure 4: Topic about presidential election



It is quite clear that the first topic is about Federal Reserve, second about rating agencies, third about slowdown of the economy and the last about politics or presidential election.

Interestingly, these topics seem to be very stable across different runs of the model and does correspond to real topics rather than arbitrary groups of words. Note, these topics were found to Granger couse the volume of S&P500 trading over all 5 iterations (except for the slowdown topic which did the same over first 4 iterations).

On the other hand, there were also junk topics that probably does not have any causal relationship with market variables and are spurious.

Nonetheless, it seems intuitive that the content of news precedes the volume of trading in the market. That is, the media channel basically makes market participants aware of when to take some action, but the action taken does not necessarily depend on the content of the news itself. This does not contradict findings of previous experiments - low accuracy of trying to predict states of markets using the topic mixtures within documents.

7 Conclusions

The paper investigated whether the content of news is related to market behavior. The premise behind the thesis is that the content of news could be captured (quantified) using the so called topic models and that the inferred topics (topic mixtures) is a relevant characteristic for determining the market outcomes. Intuitively, this seems reasonable: since topics in this context is distributions of words, the dynamics in topic mixtures should be a straightforward way to describe the content of news quantitatively. And so if the content of news is descriptive or causal of events that are relevant for market participants - a strong relationship should be possible to determine between the two.

The results on the other hand, does not resolve the question in a straightforward fashion. While we believe that the news data source used was representative and relevant for the financial industry and models applied did capture the content the way intended, the results somewhat disappoint. Specifically, trying to predict the return of a particular market variable at a specific period of time using inferred topic mixtures from topic models and being guided by the literature for processing the text data and selecting hyperparameters resulted in a worse accuracy when compared to ad hoc approach. In all analyzed cases, the prediction accuracy of outcomes of financial markets using text data had a marginally better result than a random guess, especially having in mind how the experiments were set up.

On the other hand, Granger causality analysis using market variable and inferred topic mixtures as time-series did not reveal any instances when

market variable would Granger cause a specific topic, while the reverse relationship was observed. In other words, it seems that the content of news does Granger cause financial market news and especially - the trading volume, at least in our experiment. While there probably have been some spurious results (some junk/unrelated topics Granger causing market variables), there were also instances of genuine and intuitively reasonable topics that seemed to be very stable across different runs of models that were Granger causing volume of S&P 500 trading.

That in mind, it seems that the content of news might not correspond exactly to how market is going to behave. On the other hand, it does look like that the media precedes the activity in the market.

To the analysis and results described above, paper also derived Direct Representation scheme for inferece for one the topic models - Hierarchical Dirichlet Process mixture model. Experiments showed, that using different representation, this model performed poorly.

On the other hand, the paper leaves place for further research. Most notably, two points should be pushed forward. Time dynamics should be incorporated into the topic models directly, because currently, topic models assume topics as being independent across documents, while in our context it is very likely, that topics in the news have some specific time dependence. Secondly, some external expertise could be use to validate the appropriateness and relatedness of inferred topics to the financial markets. Currently it is data driven based on maximizing the objective function. However, if validated by some external expertise, this could provide further confidence in the results.

8 References

[1] D. M. Blei and J. D. Laferty. Topic models, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] David. M. Blei and Jon. D. McAuliffe. Supervised topic models. In *arXiv:1003.0783*, 2010.

[4] Michael Bryant and Erik B. Sudderth. Truly nonparametric online variational inference for hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems Conference*, 2012.

[5] The Standford Natural Language Processing Groupd. Stemming and lemmatization. http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[6] S. Hansen, M. McMahon, and A. Prat. Transparency and deliberation within the fomc: A computational linguistics approach. In *CEP Discussion Paper No 1276*, 2014.

[7] M. C. Hughes, D. I. Kim, and E. B. Sudderth. Reliable and scalable variational inference for the hierarchical dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, 2015.

[8] M. C. Hughes and E. B. Sudderth. Memoized online variational inference for dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, 2013.

[9] H.Y.Toda and T.Yamamoto. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, 66:225–250, 1995.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[11] C. Kearney and S. Liu. Textual sentiment in finance: A survey of methods and models, 2014.

[12] S. Kim and D. Kim. Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior and Organization*, 107B:708–729, 2014.

[13] M. Kula. Using textual information in econometrics: Quantifying newspaper sentiment. *Social Science Research Network*, 2012.

[14] F. Li. Do stock market investors understand the risk sentiment of corporate annual reports?, 2006.

[15] W. Medhata, A. Hassanb, and H. Korashyb. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5:1093–1113, 2014.

[16] A. K. Nassirtoussia, S. Aghabozorgia, T. Y. Waha, and D. C. L. Ngob. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41:7653–7670, 2014.

[17] R. Nyman, D. Gregory, S. Kapadia, P. Ormerod, D. Tucket, and R. Smith. News and narratives in financial systems: Exploiting big data for systemic risk assessment. preprint on webpage at risklab.fi/sra2015/papers/Nyman_et_al_paper.pdf, 2015.

[18] Stafford Philip. Symphony is just the latest attack on bloomberg's fortress, July 2015. [Online: www.ft.com/cms/s/0/cc58a080-2ed1-11e5-91ac-a5e17d9b4cff.html#axzz3ssgWKRVE; posted 20-July-2015].

[19] V. Ramiah and S. Davidson. An information-adjusted noise model: Evidence of inefficiency on the australian stock market. *Journal of Behavioural Finance*, 8:209–224, 2007.

[20] V. Ramiah, X. Xu, and I. A. Moosa. Neoclassical finance, behavioral finance and noise traders: A review and assessment of the literature, 2015.

[21] N. Shah and N.A. Smith. Predicting risk from financial reports with supervised topic models. preprint on webpage at www.cs.cmu.edu/~nasmith/papers/shah.thesis10.pdf, 2010.

[22] V. M. Strauss. Emotional values of words in finance: Anxiety about losses and excitement about gains. In *M.Sc. thesis in Social Cognition, University College Londons*, 2013.

[23] Y.W. Teh, M.I. Jordan, M.J. Beal, , and D.M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.

[24] Chong Wang, John Paisley, and David. M. Blei. Online variational inference for the hierarchical dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[25] S. Wang, L. Yu, and K. K. Lai. Crude oil price forecasting with tei@i methodology. *Journal of Systems Science and Complexity*, 18:145–166, 2005.

[26] Y. Yan, W. Xu, H. Bu, Y. Song, W. Zhang, H. Yuan, and S. Wang. Method for housing price forecasting based on tei@i methodology. *Systems Engineering*, 27:1–9, 2007.

[27] Cheng Zhang, Henrik Ek, Carl, Xavi Gratal, T. Pokorny, Florian, and Hedvig Kjellstrom. Supervised hierarchical dirichlet processes with variational inference. In *Computer Vision Workshops (ICCVW), IEEE International Conference on*, 2013.

Appendices

A Inference using Direct Representation scheme for the SHDP model

As noted in the main text, this appendix contains derivations of inference algorithm using the Direct Representation scheme for the SHDP model which follows [4], [7], [8]. Both, batch and online variational inference algorithms are developed. For review of assignment schemes of HDP see [23]. The appendix provides and includes many details for clear exposition thus some things are repeated in the main text.

A.1 Describing HDP model

The model assumes that there are two Dirichlet Processes, governing global and document level draws of atoms/topics:

$$G_0 \sim DP(\gamma,\ H)$$
$$G_j \sim DP(\alpha_0,\ G_0)$$

where $H \sim Dirichlet(\eta)$ is base (symmetric) distribution γ, α_0 are the concentration parameters and $j \in \{1, ..., D\}$, where D is the number of documents in the corpus.

Then we draw topics ψ_{jn} associated to n-th word in a document and generate words w_{jn} in the following manner:

$$\psi_{jn} \sim G_j$$
$$w_{jn} \sim Categorical(\psi_{jn})$$

where $w \in \{1, ..., W\}$, and W is the number of words in the vocabulary.

A.2 Analytic representation of HDP

However, this representation is used only for descriptive purposes. To describe the model analytically, we proceed as in [4] and use direct assignment scheme.

Top level Dirichlet process is constructed using Stick-breaking Construction:

$$\phi_k \sim H$$
$$\beta_k^{'} \sim Beta(1, \gamma)$$
$$\beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l)$$
$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

where ϕ_k can be interpreted as global topic distributions over vocabulary simplex (probabilities for specific words occurring conditionally on specific topic) and β_k - atom weights for the global Dirichlet process.

Similarly for document level processes, stick breaking principle is also used to specify G_j analytically:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

Note, same corpus levels atoms ϕ_k are used. Differently from corpus level process, atom weights $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty}$ are sampled from the Dirichlet process with base distribution $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ (the β_k are the atom weights from the corpus level stick-broken process) as in:

$$\boldsymbol{\pi}_j \sim DP(\alpha_0, \boldsymbol{\beta})$$

To complete the analytic representation, indicator variables are drawn: $z_{jn} \sim Categorical(\boldsymbol{\pi}_j)$. These variables map document and corpus level atoms:

$$\psi_{jn} = \phi_{z_{jn}}$$

so that the words can be generated as $w_{jn} \sim Categorical(\phi_{z_{jn}})$

A.3 Incorporating supervision into HDP

The supervision into the model is incorporated following [27], where the response variable y_j basically generated after the topics for document j are drawn, so it is determined by the words in document j. At first y_j is assumed to be a discrete variable with $\{1, ..., C\}$ different values. It is implemented using a soft-max (basically - multivariate logistic) function

$$p(y_j|\bar{\phi}_j, \mu) = \frac{\exp \mu_{y_j}^T \bar{\phi}_j}{\sum_{l=1}^{C} \exp \mu_l^T \bar{\phi}_j}$$

where μ will be inferred from the data. Also,

$$\bar{\phi}_j = \frac{1}{N_j} \sum_{n=1}^{N_j} \mathbf{z}_{jn}$$

where $\mathbf{z}_{jn}$ is a binary indicator vector of length K with zero components except for a component z_{jn} which is equal to 1. Intuitively, $\bar{\phi}_j$ could be thought as a cumulative topic count.

A.4 Variational inference for Supervised HDP

The goal is to estimate the posterior probability $p(\boldsymbol{w}, \boldsymbol{y}|\gamma, \alpha_o, \eta, \mu)$. Due to model complexity, variational inference is used and thus a posterior is approximated by maximizing evidence lower bound, i.e.:

$$\begin{aligned}
\log p(\boldsymbol{w}, \boldsymbol{y}|\gamma, \alpha_0, \eta, \mu) &= \log \int_{\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}} \sum_{\boldsymbol{z}} p(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}|\gamma, \alpha_0, \eta, \mu) \\
&= \log \int_{\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}} \sum_{\boldsymbol{z}} p(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}|\gamma, \alpha_0, \eta, \mu) \frac{q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z})}{q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z})} \\
&\geq \mathbb{E}_q[\log p(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}|\gamma, \alpha_0, \eta, \mu)] - \mathbb{E}_q[\log q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z})]
\end{aligned}$$

The q here represents the variational distribution which is assumed to be fully factorized which appeals to mean field variational inference :

$$q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z}) = q(\boldsymbol{\beta}) \prod_{k=1}^{\infty} q(\phi_k|\lambda_k) \prod_{j=1}^{D} q(\pi_j|\theta_j) \prod_{n=1}^{N_j} q(z_{jn}|\varphi_{jn})$$

As in earlier description, D is the number of documents in the corpus. N_j here is the number of words in document j.

Also, specific variational distributions are chosen:

$$\begin{aligned}
q(\boldsymbol{\beta}) &= \delta_{\beta *}(\boldsymbol{\beta}) \\
q(\phi_k|\lambda_k) &= Dirichlet(\phi_k|\lambda_k) \\
q(\pi_j|\theta_j) &= Dirichlet(\pi_j|\theta_j) \\
q(z_{jn}|\varphi_{jn}) &= Categorical(z_{jn}|\varphi_{jn})
\end{aligned}$$

where $\delta_{\beta^*}(\boldsymbol{\beta})$ denote degenerative distribution at point β^*. This is used because for one it simplifies derivations but also empirically it was shown to have small posterior variance [24], especially when the dataset is large.

So now, we have our variational objective function:

$$
\begin{aligned}
L(q) &:= \mathbb{E}_q[\log p(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y} | \gamma, \alpha_0, \eta, \mu)] - \mathbb{E}_q[\log q(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{z})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{w} | \boldsymbol{\phi}, \boldsymbol{z})] \\
&+ \mathbb{E}_q[\log p(\boldsymbol{y} | \boldsymbol{z}, \mu)] \\
&+ \mathbb{E}_q[\log p(\boldsymbol{\pi} | \alpha_0, \boldsymbol{\beta})] \\
&+ \mathbb{E}_q[\log \boldsymbol{z} | \boldsymbol{\pi}] \\
&+ \mathbb{E}_q[\log p(\boldsymbol{\phi} | \eta)] \\
&+ \mathbb{E}_q[\log p(\boldsymbol{\beta} | \gamma)] \\
&- \mathbb{E}_q[\log q(\boldsymbol{z} | \boldsymbol{\varphi}] \\
&- \mathbb{E}_q[\log q(\boldsymbol{\pi} | \boldsymbol{\theta}] \\
&- \mathbb{E}_q[\log q(\boldsymbol{\phi} | \boldsymbol{\lambda}]
\end{aligned}
$$

The exact expansion of each of terms is provided in the [4] except for the second term $\mathbb{E}_q[\log p(\boldsymbol{y} | \boldsymbol{z}, \mu)]$. This is because differently from [4], a supervision has been incorporated. On the other hand, differently from [27], a different construction of the model has been used which involves less additional indicator variables.

A.4.1 Update equation for φ

So since a different route is chosen compared to [27] and [4], note that for the expectation of the response term:

$$\begin{aligned}
\mathbb{E}_q[\log p(y_j|\boldsymbol{z}_j,\mu)] &= \mathbb{E}_q\left[\log \frac{\exp\left(\mu_{y_j}^T(\frac{1}{N_j}\sum_{n=1}^{N_j}\mathbf{z}_{jn})\right)}{\sum_{l=1}^{C}\exp\left(\mu_l^T(\frac{1}{N_j}\sum_{n=1}^{N_j}\mathbf{z}_{jn})\right)}\right] \\
&= \mu_{y_j}^T\left(\frac{1}{N_j}\sum_{n=1}^{N_j}\mathbb{E}_q[\mathbf{z}_{jn}]\right) - \mathbb{E}_q\left[\log\sum_{l=1}^{C}\exp\left(\mu_l^T(\frac{1}{N_j}\sum_{n=1}^{N_j}\mathbf{z}_{jn})\right)\right] \\
&= \mu_{y_j}^T\left(\frac{1}{N_j}\sum_{n=1}^{N_j}\varphi_{jn}\right) - \mathbb{E}_q\left[\log\sum_{l=1}^{C}\exp\left(\mu_l^T(\frac{1}{N_j}\sum_{n=1}^{N_j}\mathbf{z}_{jn})\right)\right]
\end{aligned}$$

For the second term in the last equality, note that:

$$\begin{aligned}
\mathbb{E}_q\left[\log\sum_{l=1}^{C}\exp\left(\mu_l^T(\frac{1}{N_j}\sum_{n=1}^{N_j}\mathbf{z}_{jn})\right)\right] &\geq \log\mathbb{E}_q\left[\sum_{l=1}^{C}\prod_{n=1}^{N_j}\exp\left(\mu_l^T(\frac{1}{N_j}\mathbf{z}_{jn})\right)\right] \\
&= \log\mathbb{E}_q\left[\sum_{l=1}^{C}\prod_{n=1}^{N_j}\exp\left(\frac{1}{N_j}\prod_{k=1}^{\infty}\mu_{lk}^{[z_{jn}=k]}\right)\right] \\
&= \log\sum_{l=1}^{C}\prod_{n=1}^{N_j}\mathbb{E}_q\left[\left(\sum_{k=1}^{\infty}[z_{jn}=k]\exp\left(\frac{1}{N_j}\mu_{lk}\right)\right)\right] \\
&= \log\sum_{l=1}^{C}\prod_{n=1}^{N_j}\left(\sum_{k=1}^{\infty}\varphi_{jnk}\exp\left(\frac{1}{N_j}\mu_{lk}\right)\right)
\end{aligned}$$

So for further analysis we have updated lower bound for the response expectation term:

$$\mathbb{E}_q[\log p(y_j|\boldsymbol{z}_j,\mu)] \geq \mu_{y_j}^T\left(\frac{1}{N_j}\sum_{n=1}^{N_j}\varphi_{jn}\right) - \log\sum_{l=1}^{C}\prod_{n=1}^{N_j}\left(\sum_{k=1}^{\infty}\varphi_{jnk}\exp\left(\frac{1}{N_j}\mu_{lk}\right)\right)$$

The part of variational objective function which depends on φ is:

$$L_\varphi = \sum_{j=1}^{D}\sum_{n=1}^{N_j}\sum_{k=1}^{\infty}\varphi_{jnk}\left[\mathbb{E}_q \log \phi_{kw_{jn}} + \mathbb{E}_q \log \pi_{jk} - \log \varphi_{jnk}\right] + \sum_{j=1}^{D}\left(\mu_{y_j}^T\left(\frac{1}{N_j}\sum_{n=1}^{N_j}\varphi_{jn}\right) - \log\sum_{l=1}^{C}\prod_{n=1}^{N_j}\left(\sum_{k=1}^{\infty}\varphi_{jnk}\exp\left(\frac{1}{N_j}\mu_{lk}\right)\right)\right)$$

Problematic is the third term. Denote:

$$h_k = \log\sum_{l=1}^{C}\prod_{n=1,\ n\neq n_x}^{N_j}\left(\sum_{k=1}^{\infty}\varphi_{jnk}\exp\left(\frac{1}{N_j}\mu_{lk}\right)\right)\left(\exp\left(\frac{1}{N_j}\mu_{lk}\right)\right)$$

Then we can see that:

$$\sum_{k=1}^{\infty}h_k\varphi_{jn_xk} = \log\sum_{l=1}^{C}\prod_{n=1}^{N_j}\left(\sum_{k=1}^{\infty}\varphi_{jnk}\exp\left(\frac{1}{N_j}\mu_{lk}\right)\right)$$

Consider the inequality:

$$\log x \leq y^{-1}x + \log y - 1$$

where the equality holds if and only if $x = y$. Applying it we see that:

$$\log\sum_{k=1}^{\infty}h_k\varphi_{jnk} \leq \left(\sum_{k=1}^{\infty}h_k\varphi_{jnk}^{old}\right)^{-1}\sum_{k=1}^{\infty}h_k\varphi_{jnk} + \log\sum_{k=1}^{\infty}h_k\varphi_{jnk}^{old} - 1$$

where φ_{jnk}^{old} denotes previous value of φ_{jnk}. So that now:

$$\frac{\partial L_\varphi}{\partial\varphi_{jnk}} \geq \left[\mathbb{E}_q \log \phi_{kw_{jn}} + \mathbb{E}_q \log \pi_{jk} - \log \varphi_{jnk} - 1\right] + \left[\frac{1}{N_j}\mu_{y_jk}\right] - \left[\left(\sum_{k=1}^{\infty}h_k\varphi_{jnk}^{old}\right)^{-1}h_k\right]$$

Setting it to zero we have the update equation:

$$\varphi_{jnk} \propto \exp\left[\mathbb{E}_q \log \phi_{kw_{jn}} + \mathbb{E}_q \log \pi_{jk} + \frac{1}{N_j}\mu_{y_jk} - \left(\sum_{k=1}^{\infty}h_k\varphi_{jnk}^{old}\right)^{-1}h_k\right]$$

where basically φ_{jnk} is normalized to sum to 1.

A.4.2 Update equations for other variables

Since the incorporation of supervision affects only updates for φ, for λ and θ update are the same as in [4]:

$$\theta_{jk} \leftarrow \alpha\beta_k + \sum_{n=1}^{N_j} \varphi_{jnk}$$
$$\lambda_{kw} \leftarrow \eta + \sum_{j=1}^{D} \sum_{n=1}^{N_j} [w_{jn} = w]\varphi_{jnk}$$

Also, note: $\mathbb{E}_q[\log \phi_{kw}] = \Psi(\lambda_{kw}) - \Psi(\sum_i \lambda_{ki})$ and $\mathbb{E}_q[\log \pi_{jk}] = \Psi(\theta_{jk}) - \Psi(\sum_i \theta_{ji})$, where Ψ is the digamma function.

However, there are two parameters that do not have closed form update equations: β^* and μ_{lk}. They are updated using numeric optimization based on gradient (that is, uses partial derivatives of the objective function with respect to each variable). As a reminder, the part of the objective function that depends on β^* and μ_{lk} are respectively:

$$L_\beta = \sum_{k=1}^{K} (\alpha\beta_k - 1)\mathbb{E}[\log(\pi_{jk}] - \sum_{k=1}^{K} \log\Gamma(\alpha\beta_k) + (\gamma - 1)T_{K+1} - \sum_{k=1}^{K} \log T_k$$

where $T_k = 1 - \sum_{l=1}^{k-1} \beta_l$ and

$$L_\mu = \mu_{y_j}^T \left(\frac{1}{N_j} \sum_{n=1}^{N_j} \varphi_{jn} \right) - \log \sum_{l=1}^{C} \prod_{n=1}^{N_j} \left(\sum_{k=1}^{K} \varphi_{jnk} \exp\left(\frac{1}{N_j} \mu_{lk}\right) \right)$$

and K is the truncation level.

A.5 Online inference algorithm

Previously developed variational inference algorithm requires a full pass through the data which becomes computationally infeasable, especially in the financial domain where the text datasets can be huge. For this reason, online inference scheme is preferred which could work on chunks on data. To develop it, idea in [24] is followed:

$$L(q) = \sum_{j=1}^{D} L_j(q) = \mathbb{E}_j[DL_j(q)]$$

So now a batch of documents is randomly sampled from the corpora of size $|S|$. Document specific parameters φ_j, θ_j are updated using the data from the batch. Then, global parameters are updated given some learning rate ρ_t with $\sum_{t=0}^{\infty} \rho_t = \infty$ and $\sum_{t=0}^{\infty} \rho_t^2 \leq \infty$:

$$\begin{aligned}
\lambda_{kw} &\leftarrow (1 - \rho_t)\lambda_{kw} + \rho_t \hat{\lambda}_{kw} \\
\beta_k^* &\leftarrow (1 - \rho_t)\beta_k^* + \rho_t \hat{\beta}_k^* \\
\mu_{lk} &\leftarrow (1 - \rho_t)\mu_{lk} + \rho_t \hat{\mu}_{lk}
\end{aligned}$$

$\hat{\lambda}_{kw}$ is a set of sufficient statistics for topic k. It is estimated using previous section update equation adjusted for the fact that a batch of documents is used:

$$\hat{\lambda}_{kw} = \eta + \frac{D}{|S|} \sum_{j \in S} \sum_{n=1}^{N_j} [w_{jn} = w] \varphi_{jnk}$$

Since β_k^* and μ_{lk} cannot be updated in closed form, sufficient statistics for these parameters are estimated using gradient-based numeric optimization as previously.

A.6 Final notes

With these equation in hand, the whole inference steps are the same as in [27]: we either choose batch inference and update the parameters using full pass through the data or we choose online inference if for example the size of the data is very big and then proceed in sampled chunks.