ISSN: 2281-1346

Department of Economics and Management

DEM Working Paper Series

Assessing News Contagion in Finance

Paola Cerchiello
(Università di Pavia)

Giancarlo Nicola
(Università di Pavia)

# 139 (05-17)

Via San Felice, 5
I-27100 Pavia
http://epmq.unipv.eu/site/home.html

May 2017

Assessing news contagion in finance

P. Cerchiello*[†], G. Nicola[†]

[†] Dep. of Economics and Management Science, University of Pavia

Abstract

The analysis of news data in the financial context has gained a prominent interest in the last years. This because of the possible predictive power of such content especially in terms of associated sentiment/mood. In this paper we focus on a specific aspect of financial news analysis: how the covered topics modify according to space and time dimensions. To this purpose, we employ a modified version of topic model LDA, the so called Structural Topic Model (STM), that takes into account covariates as well. Our aim is to study the possible evolution of topics extracted from two well known news archive - Reuters and Bloomberg - and to investigate a causal effect in the diffusion of the news by means of a Granger causality test.

Our results show that both the temporal dynamics and the spatial differentiation matter in the news contagion.

JEL classification: C83, C12, E58, E61,G02, G14.
Keywords: Behavioural Finance, Financial News, Structural Topic Model, Granger causality

* paola.cerchiello@unipv.it
This version May 2017

1 Introduction and motivation

With the rapid growth of on-line information, text analysis and categorization have become core topics in many different disciplines ranging from politics to finance and all the social sciences in general. Text analytics techniques are an essential part of text mining and are used to classify documents (of any kind) and to find interesting information therein. The interpretation of text by machines, the task of natural language processing (NLP), is complex due to the richness of human language, as well as the ambiguity present at many levels, including the syntactic and semantic one. From a computational point of view, processing language means dealing with sequential, highly variable and sparse symbolic data, with surface forms that cover the deeper structures of meaning. Despite these difficulties, there are several methods available today that allow for the extraction of part of the information content present in texts. Some of these rely on handcrafted features, while others are highly data-driven and exploit statistical regularities in language. Among the statistical methods, many rely on word representations. Class based models, for example, learn classes of similar words based on distributional information, like Brown clustering [1] and Exchange clustering [2],[3]. Soft clustering methods, like Latent Semantic Analysis (LSA) [4] and Latent Dirichlet Allocation (LDA) [5], associate words to topics through a distribution over words of how likely each word is in each cluster/topic. In the last years many contributions employs machine learning and semantic vector representations [6],[7], mainly based on neural networks [8],[9], [10] to model complex and non-local relationships in the sequential input (see also [11],[12],[13] and [14]). If we focus specifically on the finance related research area, we can list several papers that take advantage of text analytics per se or as an additional source of information to be used. Central banks themself have been recently starting to recognize the utility of text data in financial risk analytics [15][16].

This recent rise of interest around text-based computational methods to be integrated for the assessment of financial risk is fuelling a rapidly growing literature that can be divided in two main streams according to the type of employed text: social media blogs and platform (namely Twitter, Facebook, Google Trends) or official news archive (above all Reuters and Blomberg).

In the first case, the constant production of detailed on-line information streaming from social networking and micro-blogging platforms, is increasingly attracting the attention of researchers and practitioners especially for the detection and monitoring of sentiments and opinions. Indeed, social media contents may constitute a relevant asset for financial institutions to gain useful insights about the clients' needs and perceptions in real time. Insofar, extracting sentiments from Twitter has been already employed for several purposes: to predict the trends of Dow Jones Index [17], to check the effects of sentiments on stock price and volume in the Dow Jones Index [18] or to predict market prices in the Italian financial market [19]. There are many other papers in this field like [20], [21], [22], [23], [24] and [25]. Another strand of literature uses social media as an alternative way to release information, thus reducing information asymmetry and improving stock liquidity, attracting more investors. Other papers such as [26] or [27] use Twitter data dynamically to see how information diffusion affects trading and how track changes in investor disagreement.

On the other hand, if we consider official news as source of information, not only sentiment but also content analysis is crucial, since the resulting outcomes are used for assessing correlation with events of interest (typically stress events). Many of the proposed approaches have been based on hand-crafted dictionaries that, despite requiring work to be adapted to single tasks, can guarantee good results due to the direct link to human emotions and the capability of generalizing well trough different datasets. Examples of this kind are the works of [28] and [29]. The first analyzes sentiment trends in news narratives in terms of excitement/anxiety and find increased consensus to reflect pre-crisis market exuberance, while the second correlates the sentiment in news with the housing market. Despite the good results, there are applications where it could be preferable to avoid dictionaries in favour of more data driven methods, which have the advantage of higher data coverage and capability of going beyond single word sentiment expression. Ref. [30] provides an example of a more sophisticated supervised corpus-based approach, in which they apply a framework modelling financial sentiment expressions by a custom data set of annotated phrases. They apply a fully data driven model with unsupervised semantic generalization, supervised only by a small set of events. In this vein, papers based on deep learning approaches have shown good results in predicting distress events of financial institutions [31, 32] and S&P500 stocks [33].

In this paper we follow this second stream of research based on official news and we deepen a particular aspect: improving information elicitation to enhance the model with contextual information (metadata, covariates) related to the characteristics and to the environment in which the entities of interest are operating. Yet, the introduction of contextual information in the models is not a straightforward process but requires a careful choice of the additional information provided in order to not increase noise. These advancements in text analytics aim at increasing the potential value of text as a source in data analysis [29]. Moreover, choosing as covariates temporal and spatial variables, will help us in understanding the possible evolution pattern or contagion effect of the information flow. In this respect, we employ a modified version of the well-known topic model LDA, called Structural Topic Model (STM), proposed by [41] in 2016 that explicitly includes covariates in the model fitting. To our knowledge, this is the first attempt to assess the contagion effect through news in finance.

The paper is organized as follows: In Section 2 we illustrate the model, in Section 3 we describe the data and the preprocessing steps, in Section 4 we present the results and in Section 5 we discuss the conclusions of the work with hints on the future developments.

2 The Model

When coping with a text analysis task, a researcher has to face several different issues ranging from the problem of polysems (multiple senses for given words) and synonyms (same meaning for different words) to the computational effort and allocation of largely sparse data matrix . One of the first effective model able to solve some of those issues is represented by Latent semantic analysis (LSA) [35]. The basic idea is to work at a semantic level by reducing the vector space through Singular Value Decomposition (SVD), producing occurrence tables that are not sparse and that help in discovering associations between documents. In order to establish a solid theoretical statistical framework in this context, in [36] a probabilistic version of LSA (pLSA) has been proposed, also known as the aspect model, rooted in the family of latent class models and based on a mixture of conditionally independent multinomial distributions for the pair words-documents. The intention from the introduction of pLSA was to offer a formal statistical framework, helping the parameters interpretation issue as well. By the way the goal was achieved only partially, in fact the multinomial mixtures, which components can be interpreted as topics, offer a probabilistic justification at words but not at documents level. In fact the latter are represented merely as list of mixing proportions derived from mixture components. Moreover, the multinomial distribution presents as many values as there are in the training documents and therefore it learns topic mixture on those trained documents. The extension to previously unseen documents is not appropriate since there can be new topics. In order to overcome the asymmetry between words and documents and to produce a real generative model, [37] proposed the LDA. The idea of such new approach emerges from the concept of exchangeability for the words in a document that unfolds in the 'bag of words' assumption: the order of words in a text is not important. In fact the LDA model is able to capture either the words or documents exchangeability unlike LSA and pLSA. On the other hand LDA is a generative model in any sense since it posits a Dirichlet distribution over documents in the corpus, while each topic is drawn from a Multinomial distribution over words. However note that [38] in 2003 have shown that LDA and pLSA are equivalent if the latter is under a uniform Dirichlet prior distribution. Obviously LDA does not solve all the issues. The main restriction embedded in LDA approach and due to the Dirichlet distribution, is the assumption of independence among topics. The immediate consequence was to tackle the issue by introducing the Correlated Topic Model (CTM), as proposed in [39]. CTM introduces correlations among topics by replacing the Dirichlet random variable with the logistic normal distribution. Unlike LDA, CTM presents a clear complication in terms of inference and parameters estimation since the logistic normal distribution and the Multinomial are not conjugate. To bypass the problem, the most recent alternative is represented by the Independent Factor Topic Models (IFTM) introduced in [40]. Such proposal makes use of latent variable model approach to detect hidden correlations among topics. The choice to explore the latent model world allows to choose among several alternatives ranging from the type of relation, linear or not linear, to the type of prior to be specified for the latent source.

In this paper we focus on one of the most recent version of the LDA model proposed by [41] in 2016. This new model called Structural Topic Model (STM) considers the explicit inclusion of covariates that can help in describing and interpreting the topics along the corpus. More specifically STM allows for covariates to influence two elements of the model: the topic prevalence and the topical content. With the former, the authors refer to the proportion of a document devoted to a topic, while the latter describes the word rates used in discussing a topic. Roberts et al. [41] take advantage of the Generalized Linear Models framework to accommodate for general covariate information (or meta-data) into topics model thanks also to two previous papers from [42] and [43].

Since STM depends upon LDA, we first summarize the latter and then we move to the former. Blei et al. in [37] defines the model as follows:

$$\begin{aligned} \theta_i &\sim Dir(\alpha), & (1)\\ \phi_k &\sim Dir(\beta), & (2)\\ z_{iw}|\theta &\sim Multinomial(\theta_i), & (3)\\ x_{iw}|z_{iw} &\sim Multinomial(\phi_{z_{iw}}) & (4) \end{aligned}$$

where d_i, $i = 1, \ldots, N$ is collection of document, x_{ij} a vector of words within each document d_i listed in a vocabulary V of size $|W|$, $w = 1, \ldots, W$, K is the number of topics with $k = 1, \ldots, K$, θ_i is the distribution of topics in document d_i, ϕ_k is the distribution of words in topic k_i and z_{iw} is the topic for w-th word in d_j.

Coming to the Structural Topic Model, [44] defines it as follows:

$$\begin{aligned} \theta_i|(X_i\gamma, \Sigma) &\sim LogisticNorm(X_i\gamma, \Sigma), & (5)\\ \phi_{ik} &\propto exp(m + k_k + k_{g_i} + k_{k_{g_i}}), & (6)\\ z_{iw}|\theta &\sim Multinomial(\theta_i), & (7)\\ x_{iw}|z_{iw} &\sim Multinomial(\phi_{iz_{iw}}) & (8) \end{aligned}$$

where $w = 1, \ldots, W$, $k = 1, \ldots, K$, X_i is the covariates matrix, γ is the coefficient vector, Σ is the covariance matrix, ϕ_{ik} is the word distribution for document d_i and k-th topic, m is a reference log-word distribution while k_k, k_{g_i} and $k_{k_{g_i}}$ are the topic group and interaction effects.

The strength of the model relies on its three different components clearly represented in the four equations of from (5) to (8) : the topic prevalence is modelled by equation (5) through a logistic normal distribution which mean is not constant but it depends on the covariate. The topical content is represented by equation 6 according to which the word occurrences is modelled in terms of log-transformed rate deviations from a corpus based distribution *m*. The parameters k_k, k_{g_i}, $k_{k_{g_i}}$ represent the specific deviations: respectively for the topic, for the covariates and for the interaction topic-covariates. Finally equations (7) and (8) comprise the central part of the model reporting the distribution of topics z_{iw} and of words x_{iw} both sampled from a Multinomial distributions. LDA and STM are similar in the core language of the model that is the sampling mechanism of the topics and of the words as appear from equations (3)-(4) and (7)-(8). The main differences is in the parameters of the Multinomials that, for the STM model, depend upon covariates.

Since our research hypothesis wants to demonstrate a contagion effect in the diffusion of topics among countries according to a temporal dimension, we need a tool to prove such effect. In the following section we introduce the Granger causality test, a well-known econometric test useful when causality is an object of interest.

2.1 Granger Causality

In a well known paper [45] Granger has proposed a useful test based on the following principle: if lagged values of X_t contribute to foresee current values of Y_t in a forecast achieved with lagged values of both X_t and Y_t, then we say *X Granger causes Y*. As was first shown by Sims [46], the Granger causality corresponds to the concept of exogeneity and it is therefore necessary to have a unidirectional causality in order to guarantee consistent estimation of distributed lag models.

In our empirical experiment we have considered the following equation:

$$y_t = \mu + \sum_{i=1}^{L} \alpha_i \cdot y_{t-i} + \sum_{i=1}^{L} \beta_i \cdot x_{t-i} + \varepsilon_t \tag{9}$$

where we want to test whether a timeseries x Granger causes the timeseries y. Our null hypothesis is therefore: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_L = 0$. Taking into account that we are dealing with monthly time series, in our tests we have considered up to three lags to take into account the effect of a financial quarter.

3 The data

In this paper we analyze two public financial news datasets from Reuters News and Bloomberg News containing respectively 106,521 and 447,145 documents. The datasets span a period from October 2006 to November 2013. Such time frame is very interesting from a financial perspective since it comprehends the sub-prime crisis started in 2007 and its following evolution with modest recovery and the beginning of the sovereign debt crisis. Moreover beside this major background topics, in this period there have been many spot hot topics which have periodically grabbed the attention of the media like for example the Madoff fraud, Barclays and Deutsche bank Libor manipulation investigation and UBS tax evasion controversy.

The datasets contain a broad variety of articles ranging from analysts' recommendations trough earning announcements to legal investigation news. All the news report the timestamp of the corresponding day. Such datasets need to be carefully inspected and cleaned according to the purpose of the analysis. In our case, the analysis focuses on the SIFIs banks (Systemically Important Financial Institution according to Basel Committee definition) and thus we cleaned the dataset in order to reduce as much as possible the non-bank related news. Then, we have tokenized each document into sentences and kept only those containing SIFI label (see table 1). We have developed a dictionary of bank names to be matched with the available sentences and we do not include bank tags and tickers due to their possible ambiguity with other entities (for example Royal Bank of Scotland's ticker RBS is also a famous Rugby Tournament). In addition, in order to associate a phrase to a single bank and to avoid multiple imputation, we have kept sentences referring only to one bank. Finally, since many of these institutions are very active in the investment banking sector and often release reports on other companies, we have dropped the sentences containing keywords associated with this kind of news, such as: ”analyst”, ”analysts”, ”said”, ”note”, ”report”, ”rating”. This selection procedure is somehow restrictive, but it is necessary to deal with a clean dataset focused only on banks related news. The phrases remaining after this filtering are 136,419 and cover many of the SIFI with the proportions reported in table 1.

Table 1: List of considered SIFI Banks

Bank	**# of sentences**	**Country**
Bank of America	19,203	USA
Goldman Sachs	16,258	USA
Citigroup	15,446	USA
UBS	13,414	Switzerland
Barclays	11,434	UK
Morgan Stanley	11,162	USA
HSBC	8,693	UK
Deutsche Bank	7,471	Germany
Credit Suisse	6,385	Switzerland
Wells Fargo	4,876	USA
Bank of China	3,416	China
Societe Generale	2,463	France
BNP Paribas	2,012	France
Royal Bank of Scotland	1,943	UK
Standard Chartered	1,813	UK
Commerzbank	1,512	Germany
BNY Mellon	1,427	USA
Credit Agricole	1,195	France
Banco Santander	1,023	Spain
State Street	926	USA
Sumitomo Mitsui	900	Japan
JP Morgan	755	USA
Industrial and Commercial Bank of China	732	China
BBVA	718	Spain
Lloyds Bank	648	UK
China Construction Bank	387	China
ING Bank	110	Netherlands
Unicredit	94	Italy
Dexia Group	2	Belgium
Total	136,418	

Regarding the covariates used in the STM model, we have considered the time stamp grouped into 80 months and a country/bank categorical variable. While the former helps us in monitoring the evolution of news along the horizon time, the latter is useful in disentangling the country/institution effect.

4 Results

To select a model with a good interpretability, we have tested different topic numbers and inspected manually the meaning of the resulting configurations. To evaluate the interpretation clarity, we have

Table 2: Distribution of documents per country

Country	**# of sentences**
USA	70,053
UK	24,531
Switzerland	19,799
Germany	8,983
France	5,670
China	4,535
Spain	1,741
Japan	900
Netherlands	110
Italy	94
Belgium	2
Total	136,418

considered the top 20 words associated to each topic according to the highest probability measure and to frequency (Frex). In [44] the FREX metric has been proposed to measure exclusivity in a way that balances word frequency. FREX is the weighted harmonic mean of the words rank in terms of exclusivity and frequency.

We tested 6 different configurations with 5, 10, 12, 15, 25, 35 (simulation time in Table 2) and we concluded that results with 10, 12 and 15 topics are consistent to each other in terms of arguments identified (see Figure 4.1).

Table 3: Topic concordance of the different STM configurations

Topic title	**10 topics**	**12 topics**	**15 topics**
UBS tax fraud scandal	✓	✓	✓
Market performance	✓	✓	✓
Stock recommendation	✓	✓	✓
Chinese company news	✓	✓	✓
Hedge Funds, Private Equity and Investment Banking	✓	✓	✓
Press comments and PR	✓	✓	✓
Citigroup bailout	✓	✓	✓
Advisory			✓
Morgan Stanley Investment Banking	✓	✓	✓
Euro area banks	✓	✓	✓
Madoff scandal			✓
Barclays and Deutsche Bank LIBOR manipulation	✓	✓	✓
Bond, Equity and CDS markets			✓
Mortgage crisis		✓	✓
Spanish banks			✓
General view on the economy		✓	

Table 4: Simulation time of the different STM configurations

# of topics	**time (s)**
5	371
10	522
12	685
15	543
25	1,155
35	6,667

To have a fair comparison, in each simulation run we applied the same data cleaning process removing English stopwords, keeping only the words with length between 4 and 15 letters appearing in more than 30 and less than 45k documents to remove both too rare and too common words. We kept also the STM model parameters set to an improvement stop criteria equal to 1e-5. In the following, we describe the 15 topics model configuration since it shows well defined and interpretable topics. Moreover, as emerges from figure 4.1 it is fully comparable to other configurations like 10 or 12 topics but with an increased level of clarity and definition and with the addition of relevant topics like 'Madoff scandal' and 'Spanish banks news'.

Our findings show that the identified topics represent some of the most discussed financial events that took place between 2007 and 2013, in particular:

'UBS tax fraud scandal' (top. 1), 'Market performance' (top. 2), 'Stock recommendation' (top. 3), 'Chinese company news' (top. 4), 'Hedge Funds, Private Equity and Investment Banking' (top. 5), 'Press comments and PR' (top. 6), 'Citigroup bailout' (top. 7), 'Advisory' (top. 8), 'Morgan Stanley Investment Banking' (top. 9), 'Euro area banks' (top. 10), 'Madoff scandal' (top. 11), 'Barclays and Deutsche Bank LIBOR manipulation' (top. 12), 'Bond, Equity and CDS markets' (top. 13), 'Mortgage crisis (top. 14), 'Spanish banks' (top. 15). For sake of completeness, we report in Table 4 the complete list of words associated to each topic according to the FREX measure that accounts for their overall frequency and how exclusive they are to the topic.

The wordcloud in figure 4.2 reports the most relevant words along the whole analysed corpus and it clearly highlights some words specifically connected to the 15 topics like Citigroup, Barclays, China, Morgan etc.

Table 5: List of 15 topics with associated words ordered according to FREX measure (words are weighted by their overall frequency and how exclusive they are to to topic

Topic	**Words**
Topic 1	FREX: charg, justic, guilti, account, ubsn, evas, plead, prosecut, crimin, hide, depart, evad, client, indict, california, avoid, wealthi, adoboli, involv, ubsnvx
Topic 2	FREX: gain, percent, cent, cmci, lost, ralli, advanc, drop, materi, sinc, jump, return, slip, tumbl, climb, slid, compil, rose, close, bloomberg
Topic 3	FREX: sumitomo, mitsui, suiss, csgn, scotland, neutral, credit, lloy, spectron, neutral, rbsl, royal, icap, mizuho, csgnvx, maker, suisse , outperform, baer
Topic 4	FREX: elec, cosco, sino, comm, lung, chem, pharm, fook, sang, shougang, yuexiu, sinotran, picc, swire, people , intl, emperor, shui, citic, hang
Topic 5	FREX: sach, goldman, groupinc, blankfein, sachs , gupta, rajaratnam, sachsgroup, corzin, paulson, vice, wall, rajat, tourr, presid, warren, buffett, obama, hathaway, gambl
Topic 6	FREX: spokesman, comment, charlott, spokeswoman, immedi, carolina-bas, tocom, bacn, countrywid, north, avail, lewi, moynihan, confirm, carolina, declin, respond, corp, repres, america
Topic 7	FREX: bailout, citigroup, pandit, sharehold, prefer, receiv, vikram, troubl, citigroup, announc, rescu, common, taxpay, worth, subprim, crisi, dividend, loss, plan, shed
Topic 8	FREX: advis, hire, head, team, familiar, privat, wealth, manag, appoint, deal, equiti, arrang, advisori, co-head, counsel, person, barclay, financ, dbkgnde, advic
Topic 9	FREX: stanley, morgan, stanley , smith, barney, gorman, mack, ventur, facebook, estat, bear, fuel, brokerag, underwrit, real, stearn, crude, commod, brent, healthcar
Topic 10	FREX: societ, pariba, commerzbank, euro, estim, profit, quarter, french, general, forecast, itali, greek, half, predict, germany , technic, germani, greec, socgen, incom
Topic 11	FREX: case, mellon, truste, southern, district, york, suit, bankruptci, mortgage-back, claim, stempel, oblig, collater, file, madoff, lehman, picard, jonathan, rakoff, manhattan
Topic 12	FREX: libor, manipul, diamond, regul, scandal, told, wrote, think, confer, fine, ubss, gruebel, respons, lawmak, event, england, polici, hsbcs, complianc
Topic 13	FREX: basi, point, markit, itraxx, percentag, yield, basispoint, swap, spread, preliminari, manufactur, extra, read, managers , tokyo, demand, releas, bond, econom, narrow
Topic 14	FREX: fargo, charter, chase, well, standard, jpmorgan, jpmn, home, wfcn, build, korea, portfolio, loan, francisco-bas, origin, size, mutual, small, fargo , india
Topic 15	FREX: banco, santand, bbva, bilbao, peso, spain , argentaria, spanish, chile, vizcaya, brazil, latin, mexico, spain, brasil, follow, mover, brazilian, mexican

To further evaluate topics' relevance, we report in figure 4.3 the 15 topics sorted according to their prevalence, which represents the proportion of documents devoted to each topic. Market performance, UBS scandal and Chinese news represent the most relevant and covered topics showing a prevalence greater than 0.2.

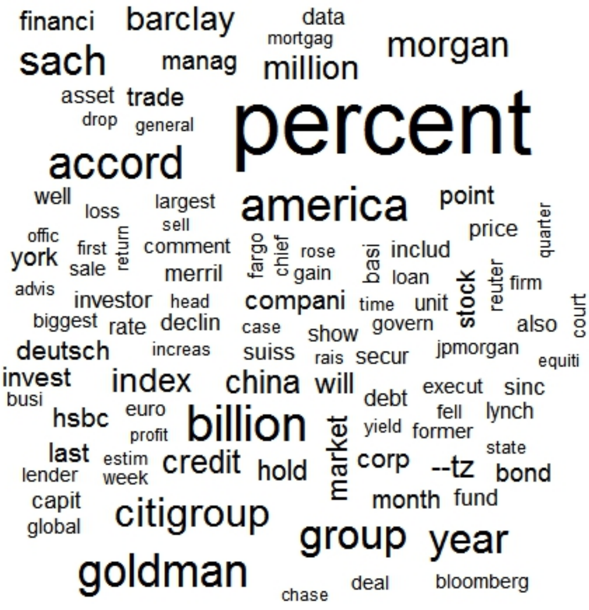


Figure 4.1: Wordcloud of the 15 topics analysis

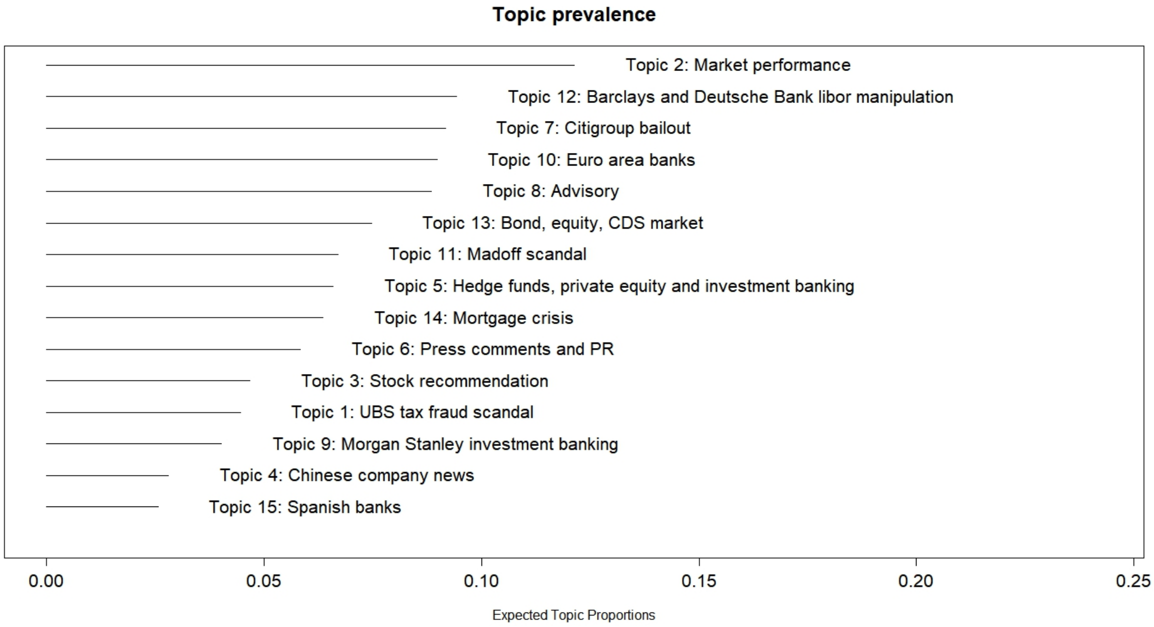
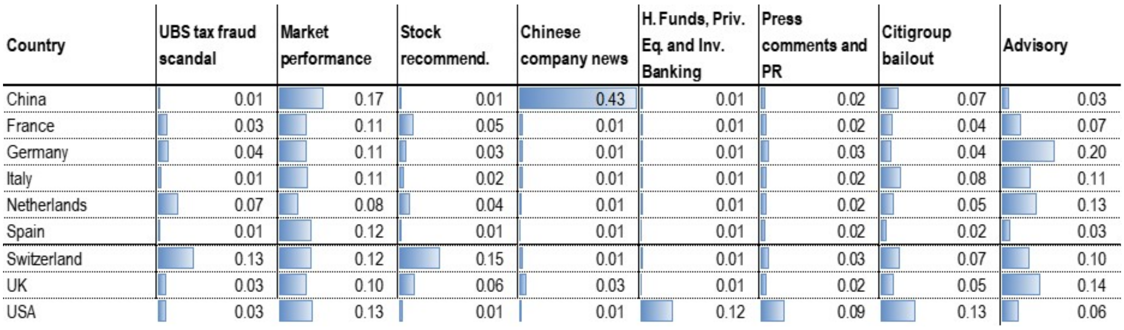


Figure 4.2: Topic prevalence the 15 topics analysis

Since the STM model allows for considering covariates that help in describing the topics, we have explicitly employed a temporal variable in terms of date of release of the news on a monthly basis and a spatial variable referring to the nationality of the bank covered within the news. Insofar, we can analyze either separately or in combination how the topics evolve through space and time. As a final aim we show the presence of a causal link in the contagion of specific topic among the analyzed countries.

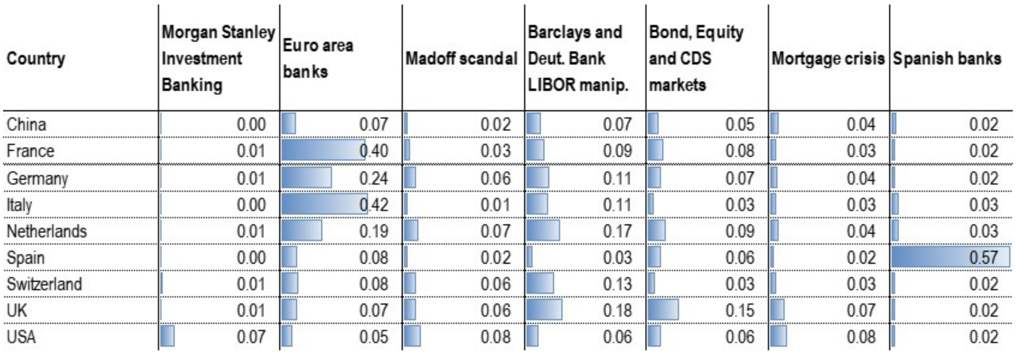
In figure 4.4 and 4.5 we map jointly the considered country and the discovered 15 topics. Such analysis allows to highlight the specific country dependence of some topics like the 'UBS scandal' upon

Switzerland, the 'Chinese news' upon China or the 'Mortgage crisis' upon USA and UK. On the other hand, we can see topics more diffused among the countries revealing a possible contagion/diffusion effect like for 'Madoff scandal', 'Libor manipulation', 'Citygroup bailout'.



Country	UBS tax fraud scandal	Market performance	Stock recommend.	Chinese company news	H. Funds, Priv. Eq. and Inv. Banking	Press comments and PR	Citigroup bailout	Advisory
China	0.01	0.17	0.01	0.43	0.01	0.02	0.07	0.03
France	0.03	0.11	0.05	0.01	0.01	0.02	0.04	0.07
Germany	0.04	0.11	0.03	0.01	0.01	0.03	0.04	0.20
Italy	0.01	0.11	0.02	0.01	0.01	0.02	0.08	0.11
Netherlands	0.07	0.08	0.04	0.01	0.01	0.02	0.05	0.13
Spain	0.01	0.12	0.01	0.01	0.01	0.02	0.02	0.03
Switzerland	0.13	0.12	0.15	0.01	0.01	0.03	0.07	0.10
UK	0.03	0.10	0.06	0.03	0.01	0.02	0.05	0.14
USA	0.03	0.13	0.01	0.01	0.12	0.09	0.13	0.06

Figure 4.3: Topic prevalence by country, topic 1-8



Country	Morgan Stanley Investment Banking	Euro area banks	Madoff scandal	Barclays and Deut. Bank LIBOR manip.	Bond, Equity and CDS markets	Mortgage crisis	Spanish banks
China	0.00	0.07	0.02	0.07	0.05	0.04	0.02
France	0.01	0.40	0.03	0.09	0.08	0.03	0.02
Germany	0.01	0.24	0.06	0.11	0.07	0.04	0.02
Italy	0.00	0.42	0.01	0.11	0.03	0.03	0.03
Netherlands	0.01	0.19	0.07	0.17	0.09	0.04	0.03
Spain	0.00	0.08	0.02	0.03	0.06	0.02	0.57
Switzerland	0.01	0.08	0.06	0.13	0.03	0.03	0.02
UK	0.01	0.07	0.06	0.18	0.15	0.07	0.02
USA	0.07	0.05	0.08	0.06	0.06	0.08	0.02

Figure 4.4: Topic prevalence by country, topic 9-15

To consider jointly the temporal and spatial effect, we further investigate some interesting topics like number 12 'Libor manipulation', number 10 'Euro area banks', number 11 'Madoff scadal' and number 14 'Mortgage crisis' that appear to be more diffused among several countries.

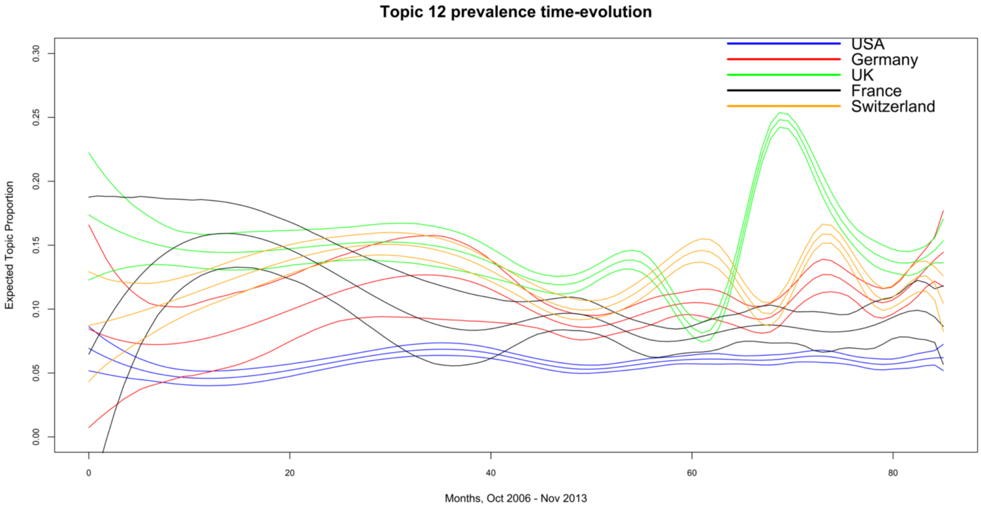


Figure 4.5: Topic evolution by country and time, topic 12

Through figure 4.6-4.9, we can have insights about a lag effects in the spikes of the news with regards to the different countries. For example in figure 4.6, related to topic 12 about Libor manipulation, it appears clearly a misalignment of the peaks specifically for UK, Switzerland and Germany, suggesting to further investigate through inferential tools. Similar considerations can be drawn for the other plots like for example figure 4.8 where the misalignment is evident for USA, Switzerland, Germany

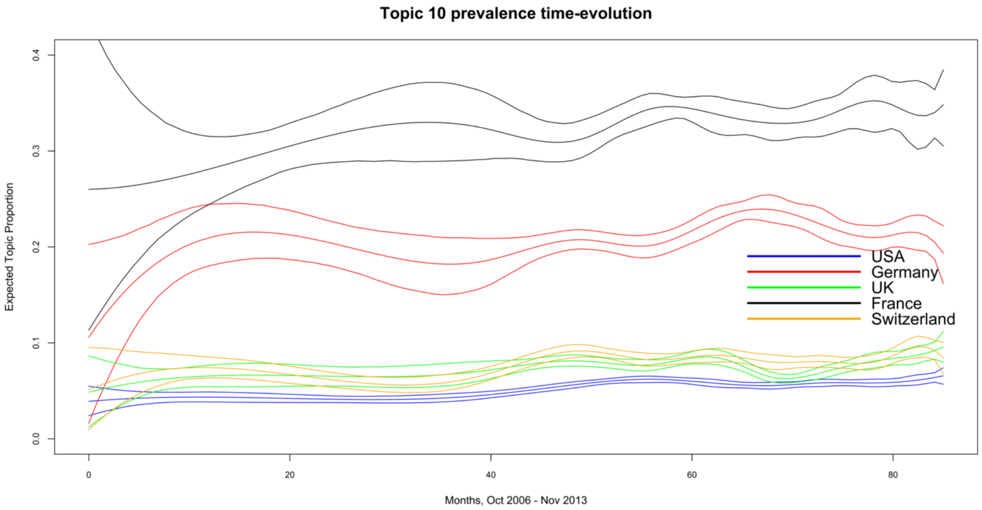


Figure 4.6: Topic evolution by country and time, topic 10

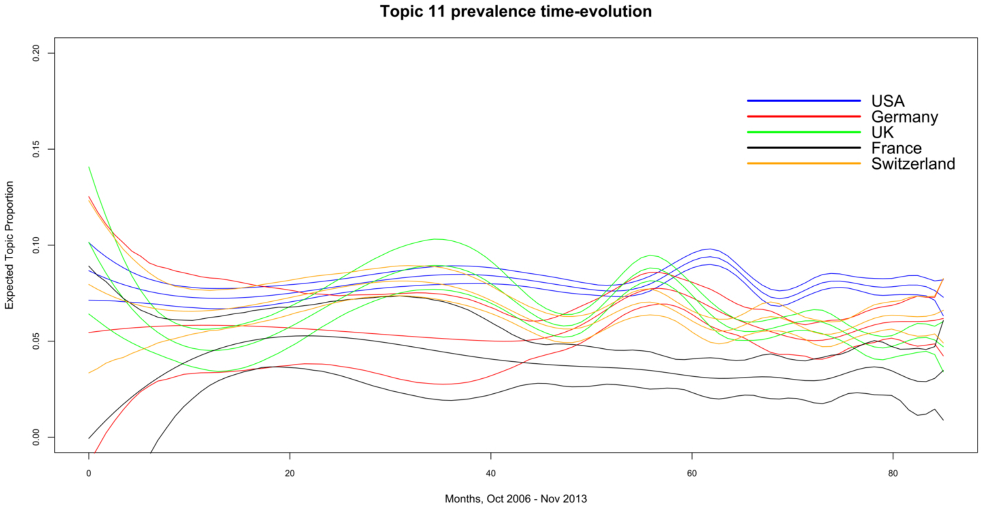


Figure 4.7: Topic evolution by country and time, topic 11

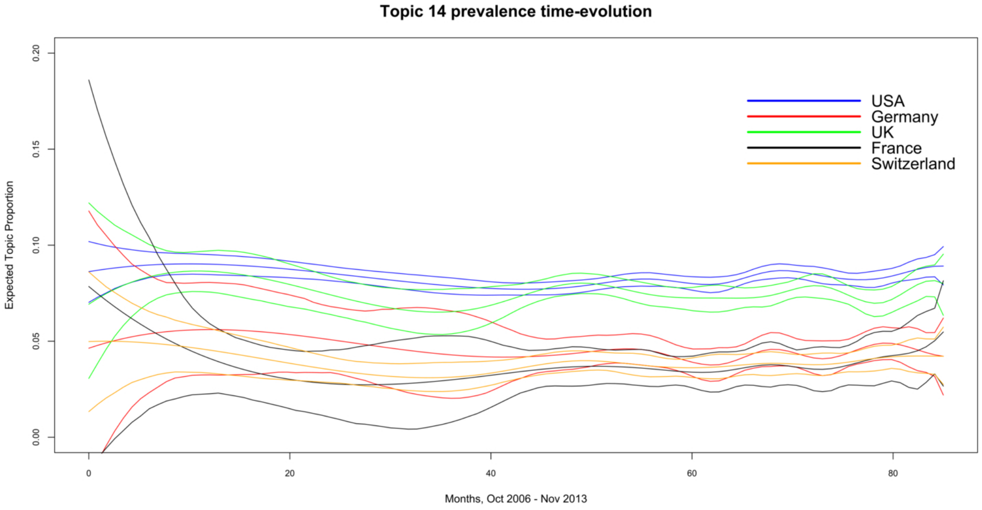


Figure 4.8: Topic evolution by country and time, topic 14

and France. To better evaluate a temporal causal effect we apply a tool particularly effective in testing such hypothesis, that is the Granger Causality test.

Among the 15 discovered topics we focus specifically on 6 arguments that we consider more important from a contagion point of view: 'UBS fraud scandal (1)', 'Citigroup bailout (7)', 'Euro area banks (10)', 'Madoff scandal (11)', 'Libor Manipulation (12)' and 'Mortgage crisis (14)'. In table 5 we report only results significant at 5% for the topics listed above, where 1L stands for 1 month lag and similarly 2L for 2 months lag. The reader can easily understand that there are several significant Granger causalities both at 1 and 2 months-lag. As one would expect, the Granger causation is both within European countries and between USA and European countries, stressing the strict interconnection among countries from a financial perspective. From this analysis we have excluded China and Japan due to a limited number of available documents that can bias results (see table 2).

Table 6: Results from Granger causality test for Topic 1-7-10-11-12-14

Topic 1	**Significant lag**	**Topic 7**	**Significant lag**
FR $\rightarrow USA$	1L, 2L	FR $\rightarrow USA$	1L, 2L
FR $\rightarrow UK$	1L, 2L	CH $\rightarrow UK$	1L, 2L
UK $\rightarrow DE$	2L	FR $\rightarrow UK$	1L
UK $\rightarrow FR$	2L	USA $\rightarrow CH$	1L, 2L
Topic 10	**Significant lag**	**Topic 11**	**Significant lag**
CH $\rightarrow USA$	1L, 2L	UK $\rightarrow USA$	1L, 2L
FR $\rightarrow USA$	1L, 2L	CH $\rightarrow USA$	1L, 2L
USA $\rightarrow UK$	1L,2L	DE $\rightarrow UK$	2L
CH $\rightarrow UK$	1L,2L	DE $\rightarrow CH$	2L
FR $\rightarrow UK$	1L,2L	FRA $\rightarrow CH$	2L
FR $\rightarrow CH$	1L,2L	-	-
FR $\rightarrow DE$	1L,2L	-	-
Topic 12	**Significant lag**	**Topic 14**	**Significant lag**
CH $\rightarrow USA$	2L	CH $\rightarrow USA$	2L
CH $\rightarrow DE$	1L	FR $\rightarrow UK$	2L
-	-	USA $\rightarrow CH$	1L, 2L
-	-	FR $\rightarrow CH$	2L
-	-	USA $\rightarrow FR$	1L, 2L
-	-	USA $\rightarrow DE$	1L

As example, let us focus on results for topic 11 (Madoff scandal) and topic 14 (Mortgage crisis). Regarding the former, we can see that the influencing countries at 1 month lag are UK and CH whose banks had a high exposition towards the fraud, in particular HSBC, RBS and UBS. The importance of these two countries in the topic is justified from the fact that we are considering only banks' related news focusing primarily on the relation between banks and the fraud and thus on the most exposed banks. In the Mortgage crisis we can see how the information contagion is transmitted from USA to some European countries at 1 month lag, namely FR, DE and CH, and this is a plausible result as this specific financial crisis had origin in United States. It is also interesting to pose attention to topic 10 regarding Euro area banks. All the interactions are significant both at 1 and 2 lag, and France seems to play a key role in spreading the topic among all the other European countries and also in the case of USA.

5 Concluding Remarks

In this work we have presented a fully data-driven methodology for the evaluation of news contagion through space and time. We focused on SIFIs related news taken from two public dataset from

Reuters News and Bloomberg News containing in total 553,666 documents spanning a period from October 2006 to November 2013. The aim of this study is to propose an approach for assessing the spread of news contagious among countries. To this purpose we have employed a model for topic modelling, called STM, able to fit the best topic distribution on the basis of useful covariates that can be chosen by the analyst. The introduction of time and country specific variables has allowed us to add a temporal and spatial dimension to the analysis. This information have been exploited to investigate the dynamic of news spread among countries.

In particular, we have used the Granger causality test to demonstrate a contagion/causation dynamic in the diffusion of the news employing times series counts extrapolated from the STM approach. Results are promising, we have found several significant causal relations in the diffusion of the news, stimulating further development in a future work. In particular, we shall investigate a correlation structure in the news diffusion taking into account county or bank level with correlation network models. Moreover the analysis should be conducted at an higher level of granularity that is at least with weekly based data. However, for some country/bank combinations, this would mean a not sufficient data coverage, possibly producing a bias in the results.

References

[1] Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., and Mercer, R. L.. Class-based n-gram models of natural language. Computational Linguistics, 18(4):467479, 1992.

[2] Martin, S., Liermann, J., and Ney, H.. Algorithms for bigram and trigram word clustering. Speech Communication, 24, 1937, 1998.

[3] A. Clark. Combining distributional and morphological information for part of speech induction. In Proc. of EACL, 2003.

[4] Landauer, T.; Foltz, P. W.; Laham, D.. Introduction to Latent Semantic Analysis. Discourse Processes. 25 (23): 259284, 1998. doi:10.1080/01638539809545028

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. JMLR, 3:9931022, 2003.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Proceedings of Workshop at International Conference on Learning Representations, 2013.

[7] J. Pennington, R. Socher, C. D. Manning. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 15321543, October 25-29, 2014, Doha, Qatar.

[8] S. Hochreiter, J. Schmidhuber. Long Short-Term Memory. Neural Computation 9(8):1735-1780, 1997

[9] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP, 2014.

[10] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP 2013, pages 16311642.

[11] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and Christopher D. Manning. 2011. SemiSupervised Recursive Autoencoders for Predicting Sentiment Distributions. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[12] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 16311642, Stroudsburg, PA, October. Association for Computational Linguistics.

[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12:2493- 2537, 2011.

[14] N. Kalchbrenner, E. Grefenstette, P. Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In Proceedings of ACL 2014

[15] D. Bholat, S. Hansen, P. Santos, and C. Schonhardt-Bailey. Text mining for central banks. In Centre for Central Banking Studies Handbook, volume 33. Bank of England, 2015.

[16] J. Hokkanen, T. Jacobson, C. Skingsley, and M. Tibblin. The Riksbanks future information supply in light of Big Data. In Economic Commentaries, volume 17. Sveriges Riksbank, 2015.

[17] Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2, 1-8.

[18] Ranco, G., D. Aleksovski, G. Caldarelli, M. Grcar, and I. Mozetic (2015). The Effects of Twitter Sentiment on Stock Price Returns. Plos one 1, 121.

[19] Cerchiello, P. and P. Giudici (2015). How to measure the quality of financial tweets. Quality Quantity, 119.

[20] Sprenger, Timm O. and Welpe, Isabell M., Tweets and Trades: The Information Content of Stock Microblogs (November 1, 2010). Available at SSRN: https://ssrn.com/abstract=1702854 or http://dx.doi.org/10.2139/ssrn.1702854

[21] Brown, E.D., 2012. Will Twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. In: Proceedings of the Southern Association for Information Systems Conference. Atlanta: SAIS. pp.36-42.

[22] Mittal, A. and Goel, A., 2012. Stock prediction using Twitter sentiment analysis. Working Paper, Stanford University CS 229.

[23] Rao, T. and S. Srivastava (2012). Twitter sentiment analysis: How to hedge your bets in the stock markets. CoRR abs/1212.1107.

[24] Nann, Stefan; Krauss, Jonas; and Schoder, Detlef, ”Predictive Analytics On Public Data - The Case Of Stock Markets” (2013). ECIS 2013 Completed Research. 102.

[25] Oliveira, N., Cortez, P., and Areal, N., 2013. On the predictability of stock market behaviour using stock twits sentiment and posting volume. Progress in Artificial Intelligence, Lecture Notes in Computer Science, 8154, pp.355-365. http://dx.doi.org/10.1007/978-3-642-40669-$0_3$1

[26] Chawla, N., Z. Da, J. Xu, and M. Ye (2016). Working paper - university of notre-dame. Information Diffusion on Social Media: Does It Affect Trading, Return, and Liquidity?.

[27] Giannini, R., Irvine, P., Shu, T., 2013. Do local investors know more? A direct examination of individual investors information set. Working paper. BlueCrest Capital Management, Texas Christian University, and University of Georgia.

[28] R. Nyman, D. Gregory, K. Kapadia, P. Ormerod, D. Tuckett, and R. Smith. News and narratives in financial systems: exploiting big data for systemic risk assessment. BoE, mimeo, 2015.

[29] C. K. Soo. Quantifying animal spirits: news media and sentiment in the housing market. Ross School of Business Paper No. 1200, 2013.

[30] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4):782796, 2014.

[31] S. Rönnqvist, P. Sarlin. Bank distress in the news: Describing events through deep learning. Forthcoming in Neurocomputing 2017.

[32] P. Cerchiello, G. Nicola, S. Rönnqvist, P. Sarlin. Deep learning bank distress from news and numerical financial data, Working paper, Submitted.

[33] X. Ding, Y. Zhang, T. Liu, J. Duan. Deep Learning for Event-Driven Stock Prediction. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015). 2015.

[34] P. Cerchiello, P. Giudici and G. Nicola. 2017. Twitter data models for bank risk contagion. Forthcoming in Neurocomputing.

[35] S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, Vol. 41, No. 6, 1990, pp. 391-407. doi:10.1002/(SICI)1097-4571(199009)41:6¡391::AID-AS I1¿3.0.CO;2-9

[36] T. Hofmann, Probabilistic Latent Semantic Indexing, Proceedings of Special Interest Group on Information Retrieval, New York, 1999, pp. 50-57.

[37] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. JMLR, 3:9931022, 2003.

[38] M. Girolami and A. Kaban, On an Equivalence between PLSI and LDA, Proceedings of Special Interest Group on Information Retrieval, New York, 2003, pp. 433-434.

[39] D. M. Blei and J. D. Lafferty, Correlated Topic Models, Advances in Neural Information Processing Systems, Vol. 18, 2006, pp. 1-47.

[40] D. Putthividhya, H. T. Attias and S. S. Nagarajan, Independent Factor Topic Models, Proceeding of International Conference on Machine Learning, New York, 2009, pp. 833-840.

[41] Roberts, M. E., Stewart, B. M., and Tingley, D. Navigating the Local Modes of Big Data: The Case of Topic Models" In Data Analytics in Social Science, Government, and Industry. New York: Cambridge University Press. 2016.

[42] Mimno, David, and Andrew McCallum. "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression." UAI, 2008.

[43] Eisenstein, Jacob, Amr Ahmed, and Eric P. Xing. "Sparse additive generative models of text." (2011).

[44] Roberts ME, Stewart BM, Airoldi E (2016b). A model of text for experimentation in the social sciences. Journal of the American Statistical Association, 111(515), 9881003.

[45] Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". Econometrica. 37 (3): 424438. doi:10.2307/1912791. JSTOR 1912791.

[46] Sims, C. (1972). Money, Income and Causality. American Economic Review 62, 540552.