Bank distress in the news:
Describing events through deep learning

Samuel Rönnqvist[1,2] and Peter Sarlin[3,4]

[1] Turku Centre for Computer Science – TUCS
Department of Information Technologies,
Åbo Akademi University, Turku, Finland
`sronnqvi@abo.fi`[*]

[2] Applied Computational Linguistics Lab
Goethe University Frankfurt am Main, Germany

[3] Department of Economics
Hanken School of Economics, Helsinki, Finland

[4] RiskLab Finland
Arcada University of Applied Sciences, Helsinki, Finland
`peter@risklab.fi`

Abstract. While many models are purposed for detecting the occurrence of significant events in financial systems, the task of providing qualitative detail on the developments is not usually as well automated. We present a deep learning approach for detecting relevant discussion in text and extracting natural language descriptions of events. Supervised by only a small set of event information, comprising entity names and dates, the model is leveraged by unsupervised learning of semantic vector representations on extensive text data. We demonstrate applicability to the study of financial risk based on news (6.6M articles), particularly bank distress and government interventions (243 events), where indices can signal the level of bank-stress-related reporting at the entity level, or aggregated at national or European level, while being coupled with explanations. Thus, we exemplify how text, as timely, widely available and descriptive data, can serve as a useful complementary source of information for financial and systemic risk analytics.

1 Introduction

Text analytics presents both major opportunities and challenges. On the one hand, text data is rich in information and can be harnessed in traditional ways such as for prediction tasks, while its descriptive depth also supports qualitative and exploratory, yet highly data-driven, analysis. On the other hand, decoding and utilizing the expressive detail of human language is prohibitively difficult.

---

[*] Corresponding author

In computational terms, text consists of high-dimensional and often ambiguous symbolic input (words), the semantics of which is a product of complex interactions between parts of the sequences in which they occur (phrases, sentences, paragraphs, etc.). Text is referred to as sparse data due to the high variability relative to number of samples, and unstructured data as the underlying linguistic structure must be inferred from the surface form as part of the analysis process.

We recognize that many applications of text analytics use linguistically rather naïve methods, typically operating on a bag-of-words assumption, disregarding word order and operating at the symbolic word-level alone. While these applications generally constitute pioneering work in their respective areas, there is currently ample opportunity for advancement, in particular in the intersection between machine learning, computational linguistics and economics. Following the deep learning paradigm, recent developments in natural language processing [37] open up for highly data-driven but linguistically more accommodating analysis methods based on semantic representation learning, which easily can be applied to new domains and tasks.

In this paper, we propose a deep learning setup to address the challenge of building a predictive model able to detect infrequent, coinciding events based on the sparse and unstructured input of text, while leveraging the text data to describe the events as well. Our method includes a heuristic to label text by event information, unsupervised semantic modeling, predictive modeling, aggregation of the prediction signals into indices, and the eventual extraction of descriptions. The approach is to our knowledge novel in providing text descriptions of events defined by non-descriptive data. We show how it can be applied to the study of risks in the financial system, with relatively little effort required in terms of collecting data for supervision in new tasks, which can be a prohibitive aspect of text analytics.

The study of bank distress is a prime example of a field where the use of text data remains largely uncharted, typically lacking both customized linguistic resources and clear goals for how to best utilize text, which motivates the focus on adaptive methods. Supervised by only a small set of bank distress events we demonstrate that the method we put forward can provide an index over coinciding stress-related reporting in news over time, which we then use to automatically retrieve descriptions of the events. We expect the method accordingly to be applicable to any type of event that recurringly figures in text over time, in connection to specific entities.

In the following section, we discuss previous work related to the problem setting and work that has utilized text data for similar tasks. Deep learning background and our setup, including semantic modeling, predictive modeling and evaluation, extraction of descriptions and the related indices are explained in Section 3. Finally, we report our experiments in Section 4, demonstrating the applicability of our approach to the study of bank stress.

2 Related work

The automatic identification of events in chronological text such as news has been explored at least since the 90s, when a DARPA-coordinated effort was organized [1] that set the foundation for what is known as *topic detection and tracking* (TDT), where news streams are analyzed in order to identify reporting on new events as well as recurring reporting relating to earlier events. The early detection and tracking methods were data-driven, based on clustering in particular, and intended to capture any kind of event (see, e.g., [43]).

A related area of research that since has emerged, mainly stemming from the area of information extraction, is *event extraction*, which aims at extracting complex structured information about events in terms of pre-defined types of events and entities, as well as attributes of events and roles of entities (cf., e.g., [7]). The event extraction techniques focus on identifying and extracting more specific types of information, with explicit semantic interpretation, in contrast to the TDT approach. As the information of interest is often particular to an expert domain and task, the techniques tend to require substantial expert guidance in terms of designing linguistic patterns or annotating text, which makes them less applicable in new domains where fewer resources may be available to target specific tasks and the information of interest may be difficult to strictly define. Efforts focusing on the financial domain and identification of specific types of risk include [8,9] and [16]. Tanev et al. [40] also explore the combination of data-driven preprocessing with the knowledge-driven approach to extracting events, as they monitor violent and disaster events in news. Hogenboom et al. [17] provide a thorough overview of how event extraction has evolved in various fields.

Parallel to this view on event discovery, which naturally places description of events at its heart, non-text data sources have also been investigated for the detection of significant events, or the risk thereof, such as failure of companies using machine learning [2,12]. The focus is then primarily on estimating the likelihood that a particular type of event will occur. While the specification of events in text mining tends to be more idiosyncratic to the input data, the events in distress prediction tend to be specified by when they occur and what entities they involve, as is the case in this paper, too. Such event specifications are easier to recombine with new data, including text data given appropriate modeling.

In particular, prediction of bank distress has been a major topic both before and following the global financial crisis. Many efforts are concerned with identifying the build-up of risk at early stages, often relying upon aggregated accounting data to measure imbalances (e.g., [11,23,5]). Despite their rich information content, accounting data pose major challenges due to restricted access, as well as low reporting frequency and long publication lags. A widely available and more timely source of information is the use of market data to indicate imbalances, stress and volatility (e.g., [13,25]). Yet, market prices provide little or no descriptive information per se, and only yield information about listed companies or companies' traded instruments (such as Credit Default Swaps). This points to the potential value of text as a source for understanding events such as

bank distress. More generally, central banks are starting to recognize the utility of text data in financial risk analytics, too. [18,6]

The literature on text-based computational methods for measuring risk or distress is still rather scarce and scattered. For instance, Nyman et al. [28] analyze sentiment trends in news narratives in terms of excitement/anxiety and find increased consensus to reflect pre-crisis market exuberance, Soo [38] analyzes the connection between sentiment in news and the housing market and Cerchiello et al. [10] analyse bank risk contagion with both market prices and sentiment index. All three approaches rely on manually-crafted dictionaries of sentiment-bearing words. While such analysis can provide interesting insight as early work on processing expressions in text to study risk, the approach is generally limiting as dictionaries are cumbersome to adapt to specific tasks, incomplete and unable to handle semantics beyond single words well. Nevertheless, sentiment analysis based on such simple approaches works quite well due to the fact that it relies on human emotions as strong priors in a way that generalizes across tasks and data, and because lower recall may be countered by the scale of the data.

Malo et al. [22] explore a linguistically more sophisticated approach that models financial sentiment compositionally, although without semantic generalization, supervised by a custom data set of annotated phrases. Hogenboom et al. [16] integrate their linguistically aware event extraction techniques with the conventional Value at Risk model to account for certain cases of event-driven market effects.

Data-driven approaches, such as Wang & Hua [42] predicting volatility of company stocks from earning calls, may avoid the issues of handcrafted features and manually annotated corpora. Their method, although allegedly providing good predictive performance gains, offers only limited insight into the risk-related language of the underlying text data. It also leaves room for further improvements with regard to the semantic modeling of individual words and sequences of words, which we address. Further, Lischinsky [21] performs a crisis-related discourse analysis of corporate annual reports using standard corpus-linguistic tools, including some data-driven methods that enable exploration based on a few seed words. His analysis focuses extensively on individual words and their qualitative interpretation as part of a crisis discourse, which likewise provides rather limited insight compared to what full sentences are able to communicate. Finally, Rönnqvist & Sarlin [30] construct network models of bank interrelations based on co-occurrence in news, and assess the information centrality of individual banks with regard to the surrounding banking system, a fully data-driven approach that could be further enhanced by semantic modeling and conditioning.

In the following, we introduce the deep learning approach and our particular model, along with further relevant previous work.

3 Methods

Characterized in part by the deep, many-layered neural networks, a prevailing idea of the deep learning paradigm is that machine learning systems can become more accurate and flexible when we allow for abstract representations of data to be successively learned, rather than handcrafted through classical feature engineering. By modeling the input data before modeling specific tasks, the networks can learn about regularities in the world and generalize over them, which improves performance on supervised task learning. For a recent general survey on deep learning confer Schmidhuber [34], and for a more explicit discussion of deep learning in natural language processing see Socher & Manning [37]. Moreover, Bengio et al. [4] provide a thorough review on the emerging topic of representation learning itself.

While manually designed features help bring structure to the learning task through the knowledge they encode, they often suffer problems of being over-specified, incomplete and laborious to develop. Especially regarding natural language processing, this limits the robustness of text mining systems and their ability to generalize across languages, domains and tasks. By exploiting statistical properties of the data, features can be learned in an unsupervised fashion instead, which allows for large-scale training not limited by the scarcity of annotated data. Such intensively data-driven, deep learning approaches have in recent years led to numerous breakthroughs in application domains such as computer vision and natural language processing, where a common theme is the use of unsupervised pre-training to effectively support supervised learning of deep networks [34]. We apply the same idea in modeling event-related language in text.

3.1 Labeling text by event data

The modeling is founded on connecting two types of data, text and event data, by entities and chronology. An event data set contains information on dates and names of involved entities, relating to the specific type of event to be modeled. First, a set of regular expression patterns is used to locate the entity names as they occur in the text. Second, an event is associated by the date it occurred and by the relevant timestamp of the document.

In this paper, we focus on news text where publication date is used for matching articles in time, and entity occurrences are indexed at the sentence level. Each sentence s and occurring entity b are cross-referenced against the event data in order to cast the pair as event *coinciding* (1), *non-coinciding* (0) or ambiguous (undefined), according to an inner (W_{in}) and outer (W_{out}) time window. Formally, the label is defined as:

$$e_{s,b} = \begin{cases} 1, \text{if } d_s - d_e \in W_{in} \\ 0, \text{if } d_s - d_e \notin W_{out} \end{cases}$$

where for the intervals W holds that $W_{in} \subset W_{out}$. I.e., we label each entity occurrence and its sentence as likely to discuss the event or not likely, whereas

uncertain cases that fall outside W_{in} but within W_{out} are not used. Given this heuristic, we effectively expand the data set that is to serve as supervision signal, and the predictive model will learn to generalize across examples and associate relevant language in the text data to the modeled event type.

3.2 Modeling

We are interested in modeling the semantics of words and semantic compositionality of sequences of words to obtain suitable representations of the content of the news, to use as features for predicting events and associating text descriptions. At the word level, distributional semantics exploits the linguistic property that words of similar meaning tend to occur in similar contexts [14]. Word contexts are modeled to yield distributed representations of word semantics as vectors, as opposed to declarative formats, which allow measuring of semantic similarities and detecting analogies without supervision, given substantial amounts of text [35,36,24]. The distributional semantic modeling captures the nature of words in a broader sense, in the directions of syntax and pragmatics. These word vectors provide an embedding into a continuous semantic space where the symbolic input of words can be geometrically related to each other, thus supporting both the predictive modeling in this paper and a multitude of other natural language processing tasks (e.g., tagging, parsing and relation extraction) [37,4].

While traditionally modeled by counting of context words, predictive models have eventually taken the lead in terms of performance [3]. Neural network language models in particular have proved useful for semantic modeling, and are especially practical to incorporate into deep learning setups due to their dense vectors and the unified neural framework for learning. Mikolov et al. [24] have put forward an efficient neural method that can learn highly accurate word vectors as it can train on massive data sets in practical time (a billion words in the order of a day on standard architecture).

Subsequently, Le & Mikolov [20] extended the model in order to represent compositional semantics (cf. [26]) of sequences of words, from sentences to the length of documents, which they demonstrated to provide state-of-the-art performance on sentiment analysis of movie reviews. Methods based on other neural architectures and explicit sentence structure have since gained slightly improved performance [19,39], but require parse trees as pre-structured input and are therefore not as flexible. Analogous to the sentiment analysis task, we employ the distributed memory method of Le & Mikolov to learn vectors for sentences in news articles, where entities are mentioned, and use them for learning to predict the probability of an event. Hence, when providing bank distress events, the task can be understood as a type of risk sentiment analysis that models language specific to the type of event, rather than more general expression of emotions explicitly.

Text sequences are also commonly modeled by recurrent neural networks such as Long Short-Term Memory (LSTM) networks [15], but these are not as efficient as feed-forward topologies with fixed context size in terms of speed. The sentence vector we use is a practical fixed-size representation suitable as input to
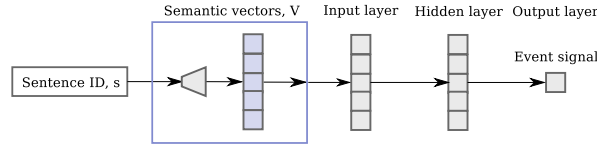
a) Semantic pre-training
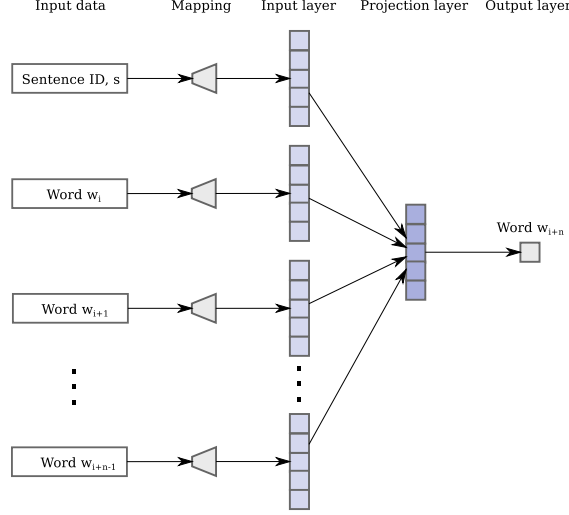
b) Supervised training

Fig. 1. Deep neural network setup for (a) pre-training of semantic vectors, and (b) supervised training against event signal e.

a feed-forward network. The input sequence of words may have a vocabulary size in the order of a million words, but the sentence vector represents the necessary semantics of each sentence as a single dense vector with a dimensionality of typically 50-1000. The reduction from sparse sequence to a fixed-length, dense representation helps train the predictive model against a signal corresponding to a comparatively tiny number of events.

Our deep neural network for predicting events from text, outlined in Fig. 1, is trained in two steps: through learning of sentence vectors as pre-training (1a), followed by supervised learning against the event signal e (1b). The use of the distributed memory model of [20] in the first step is explained in the following.

The modeling of word-level semantics works by running a sliding window over text, taking a sequence of words as input and learning to predict the next word (e.g., the 8th in a sequence), using a feed-forward topology where a projection layer in the middle provides the semantic vectors once the connection weights have been learned. A semantic vector V_i is the fixed-length, real-valued pattern

of activations reaching the projection layer for network input i. The projection layer provides a linear combination that enables efficient training on large data sets, which is important in achieving accurate semantic vectors. In addition, the procedure of [20] for sentence vector training includes the sentence ID as input, functioning as a memory for the model that allows the vector to capture the semantics of continuous sequences rather than only single words; the sentence ID in fact can be thought of as an extra word representing the sentence as global context and informing the prediction of the next word. While the prediction from word context to word constitutes a basic neural language model, the sentence ID conditions the model on the sentence and forces the sentence vector to capture the semantics that is particular to the sentence rather than the language overall. Formally, the pre-training step seeks to maximize the average log probability:

$$\frac{1}{k-n} \sum_{i=1}^{k-n} \log p(w_{i+n}|s, w_i, ..., w_{i+n-1})$$

over the sequence of training words $w_1, w_2, ..., w_k$ in sentence s with word context size n. In the neural network, an efficient binary Huffman tree coding is used to map sentence IDs and words to activation patterns in the input layer and the hierarchical softmax output layer (by referencing vectors of a matrix D), which imposes a basic organization of words by frequency. The projection layer output is a function of the average of sentence vector V_s and word vectors of the context $\{V_{w_j}|j \in [i, i+n[\}$, which means that a single vector can easily be extracted once the model is trained. The sentence vector is extractable as $V_s = \beta + UD_s$, where U is the learned projection layer weight matrix and β is the bias parameter.

The second modeling step (Fig. 1b) is a normal feed-forward network fed by the sentence vectors V_s (pertaining to the set of sentences S), which we train by Nesterov's Accelerated Gradient [27] and backpropagation [32] to predict distress events $e \in \{0, 1\}$. Hence, the objective is to maximize the average log probability:

$$\frac{1}{|S|} \sum_{s \in S} \log p(e_s|V_s)$$

The network has two output nodes for $e \in \{0, 1\}$ in a softmax layer that applies a cross-entropy loss function. In the trained network, the posterior probability $M(V_s) = p(e_s = 1|V_s)$ reflects the relevance of sentence s to the modeled event type and is derived by:

$$p(e_s = j|V_s) = \frac{e^{y_j}}{e^{y_0} + e^{y_1}}; \; y = \sigma(\beta^2 + U^2\sigma(\beta^1 + U^1V_s))$$

where σ can be any non-linear activation function (e.g., sigmoid, hyperbolic tangent or rectified linear) and U are again the learned weight matrices.

In the following sections, we discuss how the model is used for classification and evaluated by its classification performance, as we apply a threshold on the model output $M(V_s)$, as well as on aggregate functions of it.

3.3 Evaluation and aggregation

Assuming that the distribution of events for a particular entity is sparse over time, the procedure for matching events to text produces examples with skewed class frequencies. Moreover, it is likely that the user has an imbalanced preference between types of errors, preferring a sensitive system to detect possible events and provide means for further investigation in the form of descriptions, rather than missing an event. This requires extra care in evaluation.

We evaluate the performance of the predictive model to guide hyperparameter optimization and asses the quality of indices that it will produce, and importantly to provide a quantitative quality assurance for the information content of the descriptions we extract. We use the relative Usefulness measure (U_r) by Sarlin [33], as it is commonly used in distress prediction and intuitively incorporates both error type preference (μ) and relative performance gain of the model over consistently choosing the majority class. Based on the combination of negative/positive observations (obs $\in \{0, 1\}$) and negative/positive predictions (pred $\in \{0, 1\}$), we obtain the cases of true negative ($TN \equiv obs = 0 \wedge pred = 0$), false negative ($FN \equiv obs = 1 \wedge pred = 0$), false positive ($FP \equiv obs = 0 \wedge pred = 1$) and true positive ($TP \equiv obs = 1 \wedge pred = 1$), for which we can estimate probabilities when evaluating our predictive model. Further, we define the baseline loss L_b to be the best guess according to prior probabilities $p(\text{obs})$ and error preferences μ (Eq. 1) and the model loss L_m (Eq. 2):

$$
L_b = \min \begin{cases} \mu \cdot p(\text{obs} = 1) \\ (1 - \mu) \cdot p(\text{obs} = 0) \end{cases} \tag{1}
$$

$$
L_m = \mu \cdot p(FN) + (1 - \mu) \cdot p(FP) \tag{2}
$$

From the loss functions we derive Usefulness in absolute (U_a) and relative terms (U_r):

$$
U_r = \frac{U_a}{L_b} = \frac{L_b - L_m}{L_b} \tag{3}
$$

While absolute Usefulness U_a measures the gain vis-à-vis the baseline case, relative Usefulness U_r relates gain to that of a perfect model (i.e., Eq. 5 with $L_m = 0 \Rightarrow U_a = L_b$). Usefulness functions both as a proxy for benchmarking the model (testing) and to optimize its hyperparameters (validation). Usefulness can also be related to the in text mining widely used F-score[41] (based on precision $= p(\text{obs} = 1 | \text{pred} = 1)$ and recall $= p(\text{pred} = 1 | \text{obs} = 1)$):

$$
F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{4}
$$

which similarly can account for varying preferences by its β parameter, although not gain. The F_β-score assigns β times as much importance to recall as to precision (i.e., preference for completeness over exactness)[41], which is analogous to but not directly transferable to the μ parameter in the Usefulness measure.

While the F-score is commonly seen to maximize completeness versus exactness of true positives, the parameter can also be seen as a priority to minimize false negatives versus false positives (FN prioritized over FP when $\beta > 1$). As a heuristic, we map the balanced, standard F_1-score with $\beta = 1$ to U_r with $\mu = 0.5$, and match deviations from these preferences according to $\beta = \mu/(1-\mu)$.

In order to influence the sensitivity of the model, we may classify a sentence by a threshold on the positive-class posterior probability:

$$p(e_s = 1|V_s) \geq t$$

The threshold is optimized on the validation set with respect to Usefulness at a given preference, and applied to the test set for evaluation.

However, evaluating classification at an aggregated entity level rather than the level of sentence instances is more suitable to the use case, and likely more robust as the classification then combines evidence from multiple observed occurrences in the text. Instead of the direct posterior probability, at the entity level we classify by the index defined in Eq. 5 below; i.e., an event is signaled for the entity if:

$$I(p, b) \geq t$$

Furthermore, evaluation on the sentence vector level with a randomized set split into train, validation and test set may produce somewhat optimistic results, as specific language related to one particular event can be expected to be shared among several instances. Thus, the evaluation would not truly reflect how well the model can be expected to generalize across events of the same type, including future occurrences. To counter the bias, we sample the cross-validation folds according to a *leave-N-entities-out* strategy (or leave-N-banks-out), based on entity rather than sentence instance, such that discussion about a particular entity is compartmentalized into a single set. In case of very frequent entities that would cause very skewed fold sizes, the instances may be split by period such that the more recent occurrences are placed in the latter set (e.g., test rather than validation set) to minimize possible cross-contamination.

3.4 Event indices

By aggregating posterior probabilities we form an index to reflect the level of event-related reporting about an entity over time, thereby guiding exploration and extraction of descriptions, while it also serves as the signal that we evaluate against. The entity-level relevance index $I : p \times b \to [0,1]$ is formalized as:

$$I(p, b) = \frac{1}{|S_{p,b}|} \sum_{s \in S_{p,b}} M(V_s) \tag{5}$$

over the sentences $S_{p,b}$ that mention entity b in period p, where $M(V_s) = p(e_s = 1|V_s)$ gives the posterior probability of the trained neural network model.

In order to obtain better overview, it is motivated to further group entities and aggregate their indices. In the experiments, we first aggregate from sentences

to banks, and then from banks to countries to highlight national differences across Europe. The second-level index (or country-level index) is a weighted average, defined as:

$$I'(p, c) = \frac{1}{|B_c|} \sum_{b \in B_c} I(p, b) \cdot |S_{p,b}|$$ (6)

where B_c is the set of entities in category/country c. Finally, we define a top-level index that summarizes the level of relevant reporting for all modeled entities as a global average of vectors:

$$I''(p) = \frac{1}{|S_p|} \sum_{s \in S_p} M(V_s)$$ (7)

where S_p is the set of vectors for all entity-mentioning sentences in period p.

3.5   Extraction of descriptions

As the neural network in the second step of the setup has been trained and the hyperparameters optimized by cross-validation, it can be applied to sentence vectors V in order to use the posterior probability $M(V)$ as a relevance score with respect to the event type. The indices (Eq. 5, 6 and 7) provide overview over time and can highlight peaks and periods with elevated volumes of event-related discussion, which can be more closely investigated by retrieving descriptions of the underlying events.

Given a specific period and entity or set of entities, the basic principle in retrieving descriptions is to filter and rank pieces of text based on the posterior probability of the predictive model for the corresponding semantic vector. In the current setup, we perform the semantic modeling on the sentence level, which simplifies the process of retrieving relevant and specific passages. The semantic modeling can be applied to any type of textual unit, including complete documents, but that requires additional measures for locating the interesting parts within the broader context. Rönnqvist and Sarlin [31] explore this by similarly training a predictive model on document vectors and successfully applying it on word vectors, to weight the relevance individual words within the context. In current experiments, we find that, while their method works for document vectors that are trained on a larger number of words per vector, it does not work as well for sentence vectors, as they tend to be less similar to the word vectors of the same model. Overall, the extracts as presented in Section 4 are qualitatively better when produced based on sentence vectors.

Vectors are trained only for sentences that mention target entity names, as it would be infeasible in terms of memory to model each sentence separately for a large corpus, and because the direct discussion about the entities is of primary interest. The near context of such sentences however tend to support interpretation and are useful to include in presentation. The semantic model supports inference of vectors for at train-time unseen sentences, although with noisier results. We infer vectors and predict the relevance of the sentences immediately

before and after sentences in which entities occur, as there is strong dependency between neighboring sentences and a combined score of the expanded context may produce more robust predictions. The combined score for an excerpt is calculated as:

$$x_i = max \begin{cases} M\left(V_{S_i}\right) \\ M\left(\frac{1}{n}\sum_{j=1}^{n} V'_{S_{i-1}}\right) \\ M\left(\frac{1}{n}\sum_{j=1}^{n} V'_{S_{i+1}}\right) \end{cases} \quad (8)$$

which includes one sentence before and after sentence S_i. V' is a stochastic, inferred vector and n is the number of samples (e.g. 100).

The excerpts are ranked according to the score for presentation and offer a preview of the most prominent event-related discussion, which may be retrieved in full from the individual articles. The experiments that follow demonstrate the utility of the excerpts in highlighting the specific forces that drives the index, as we apply the method to model bank distress.

4   Experiments

We test the deep neural network setup for modeling event-related language on European bank distress events and news data, in order to demonstrate the value it can bring in helping to identify and understand past, ongoing or mounting events. In the following, we discuss the data we use, the modeling in practice, and our quantitative evaluation results. Finally, we provide a qualitative analysis of the indices and related events by means of their associated descriptions, going from the general, higher-level view to the more specific.

4.1   Data

The event data set for this study covers data on large European banks as entities, spanning periods before, during and after the global financial crisis of 2007–2009. We include 101 banks for which 243 distress events have been observed during 2007Q3–2012Q2. Following Betz et al. [5], the events include government interventions and state aid, as well as direct failures and distressed mergers. In addition, we map each bank to the country or countries where it is registered, to allow for aggregation of results to the country level.

The text data consist of news articles from Reuters online archive from the years 2007 to 2014 (Q3). The data set includes 6.6M articles (3.4B words). Bank name occurrences are located using a set of regular expressions that cover common spelling variations and abbreviations. The patterns have been iteratively developed against the data to increase accuracy, with the priority of avoiding false positives (in accordance to [30]). Scanning the corpus, 262k articles and 716k sentences are found to mention any of the 101 target banks.

We set the inner time window from 8 days before to 45 days after the event ($W_{in} = [-8, 45]$), and the outer from 120 days before to 120 days after ($W_{out} = [-120, 120]$), as optimized through the evaluation scheme discussed in Section

3.3. In total, 386k sentences are successfully labled and used for training and evaluation, as they fall within the span of the event data and are not deemed ambiguous cases. As expected, the class distribution is highly skewed, with 9.0% of the 386k cases being labeled as coinciding.

4.2 Semantic pre-training

First, the semantic pre-training step is performed to obtain sentence vectors for each of the 716k sentences, to be used both for training, evaluation and deployment of the model. In order to improve the word representations of the model, by extending the data coverage and letting them capture the semantics of both general English in news reporting as well as bank-specific language, the rest of the corpus is also sampled. This is achieved by running the model without the sentence-ID-related component for sentences without bank occurrences. The whole training process is repeated in multiple iterations with decreasing learning rate. We optimized the sentence vector length to 600 and context size to 5 by cross-validation. We also tested the influence of text sequence lengths, and found that training a vector on multiple sentences achieved slightly worse predictive performance, while vectors trained at sentence and document level were comparable.

4.3 Predictive modeling and evaluation

Following the semantic pre-training, we train a predictive neural network model with 3 layers. The input layer has 600 nodes, corresponding to the semantic vectors, and the output layer has two nodes corresponding to distress/tranquil states. A set of tuples including sentence vectors V_s, entity b and labels $e_{s,b}$ are complied as data for modeling.

We evaluate the predictive model with the four combinations of sampling method and level of evaluation, discussed in Section 3.3. The baseline evaluation with random sampling at the level of sentence vectors is reported in Table 1 (left), providing 27.5% relative Usefulness, i.e., performing significantly better than majority class prediction even with the highly skewed class distributions. By comparison, evaluation at the aggregated bank level (classifying by $I(p, b)$ (Eq. 5) rather than $M(V)$) reduces noise from single sentences and stabilizes prediction, thereby increasing performance to 32.6% (Table 1, center). These results show that the model is effective in linking the relevant pieces of text to the bank distress events, hence, providing a first assurance of the quality of the descriptions we will retrieve. Further, we evaluate based on leave-N-banks-out sampling, i.e., the cross-validation folds of vectors are organized by bank, such that the vectors of banks used for testing are held out of training. While this produces lower Usefulness scores, it is a more realistic estimate of future performance in the context of deploying the model on unseen banks or future data. With vector-level evaluation we reach 8.3% relative Usefulness (Table 1, right), while bank-level aggregation again stabilizes prediction and improves performance to 12.3% of available Usefulness (Table 2).

μ	Random sampling				Leave-N-banks-out	
	Vector-level		Aggregated		Vector-level	
	$\bar{U}_r(\mu)$	σ_U	$\bar{U}_r(\mu)$	σ_U	$\bar{U}_r(\mu)$	σ_U
0.1	-0.004	0.004	-0.022	0.029	-0.013	0.013
0.3	-0.007	0.004	-0.015	0.013	-0.032	0.026
0.5	0.002	0.005	-0.014	0.010	-0.039	0.036
0.6	0.013	0.007	-0.015	0.012	-0.038	0.039
0.7	0.038	0.011	0.027	0.030	-0.026	0.029
0.8	0.095	0.019	0.156	0.029	-0.008	0.044
0.85	0.157	0.026	0.260	0.030	0.025	0.048
0.875	0.207	0.028	**0.326**	0.030	0.039	0.133
0.9	**0.275**	0.054	0.268	0.031	**0.083**	0.114
0.925	0.253	0.041	0.148	0.040	0.040	0.109
0.95	0.106	0.044	-0.009	0.038	-0.052	0.153

Table 1. Cross-validated predictive performance as relative Usefulness over preferences between types of error (μ), evaluated at vector and aggregated bank level with random sampling, and at vector level with leave-N-banks-out sampling.

μ	Leave-N-banks-out, aggregated							
	$\bar{U}_r(\mu)$	σ_U	$\bar{F}_\beta$	σ_F	$\bar{TN}$	$\bar{FN}$	$\bar{FP}$	$\bar{TP}$
0.1	-0.014	0.042	0.497	0.000	516	68	0	0
0.3	-0.011	0.022	0.087	0.015	516	68	0	0
0.5	-0.015	0.029	0.031	0.013	516	68	0	0
0.6	-0.013	0.027	0.032	0.020	515	68	1	0
0.7	-0.003	0.038	0.087	0.063	511	65	4	3
0.8	0.048	0.154	0.314	0.171	472	53	44	15
0.85	0.122	0.147	0.434	0.153	435	45	80	22
0.875	**0.123**	0.173	0.529	0.174	374	38	142	30
0.9	0.081	0.162	0.629	0.189	308	31	208	37
0.925	-0.006	0.173	0.741	0.190	151	14	364	54
0.95	-0.075	0.160	0.901	0.125	38	4	477	64

Table 2. Cross-validated predictive performance as relative Usefulness and F-score over preferences between types of error (μ) and recall/precision (β), evaluated at bank level with leave-N-banks-out sampling. Mean confusion matrix values are included, too.

We find the optimal network (50 rectified linear hidden nodes), hyperparameters for the NAG training algorithm to train its weights, and threshold on $M(V)$ or $I(b,p)$ for classifying $e \in \{0,1\}$, after which we evaluate performance by U_r of the optimal model. We trained the network by randomized 5-fold cross validation with one fold for validation and one for testing, in multiple reshuffles of the data set. The evaluation yielded an area under the ROC curve of 0.712 with a standard deviation $\sigma = 0.008$ with random sampling evaluated at vector level, and an area of 0.645 ($\sigma = 0.083$) with leave-N-banks-out sampling evaluated at the aggregated bank level.

Following previous studies [5,29], we make use of a skewed preference $\mu \approx 0.9$ (i.e., missing a crisis is about 9 times worse than falsely signaling one). From the viewpoint of policy, highly skewed preferences are particularly motivated when a signal leads to an internal investigation, and reputation loss or other political effects of false alarms need not be accounted for. While our model is not robust to low levels of μ, we can see in Table 2 that Usefulness is positive and peaking as μ nears 0.9. Meanwhile, F-score is reaching its maximum at the extreme preference, which is an indication of its failure to capture gain over the majority class baseline.

We conclude that at $\mu = 0.9$ with vector-level evaluation and at $\mu = 0.875$ with aggregated evaluation the model has decent predictive performance by capturing up to 33% of available Usefulness and 12% in the more conservative leave-N-banks-out sampled exercise. To relate the results we may confer Betz et al. [5] who obtain U_r of 19-42% and Peltonen et al. [29] with 58-64%. The latter incorporates network linkages, which we currently do not model, although this is possible to extract from text as well (cf. [30]). In both cases they test a selection of models to predict bank distress using conventional data sources. These are the most similar experiments available, although not necessarily strictly comparable. A direct comparison of usefulness is in principle impossible as different data and prediction tasks will yield different results, such as the broader sample and earlier forecast horizons in Betz et al. [5]. Nevertheless, our evaluation results show that we are able to extract a stress signal from text alone. While it does not surpass the performance achieved for other tasks and samples, it does achieve acceptable levels and provides a quantitative quality assurance of the text extracts. The results also point toward the likely benefit of incorporating both text and conventional data in bank distress prediction.

4.4   A descriptive stress index for Europe

Having trained the network and evaluated its predictive performance, we can reliably extract indices of stress at the different levels of aggregation together with extracts to describe them. In this section, we discuss patterns recognizable in the top-level view, with a summary of what we are able to learn from the associated descriptions. The following sections continue with a breakdown into countries and banks, which supports a more targeted qualitative analysis.
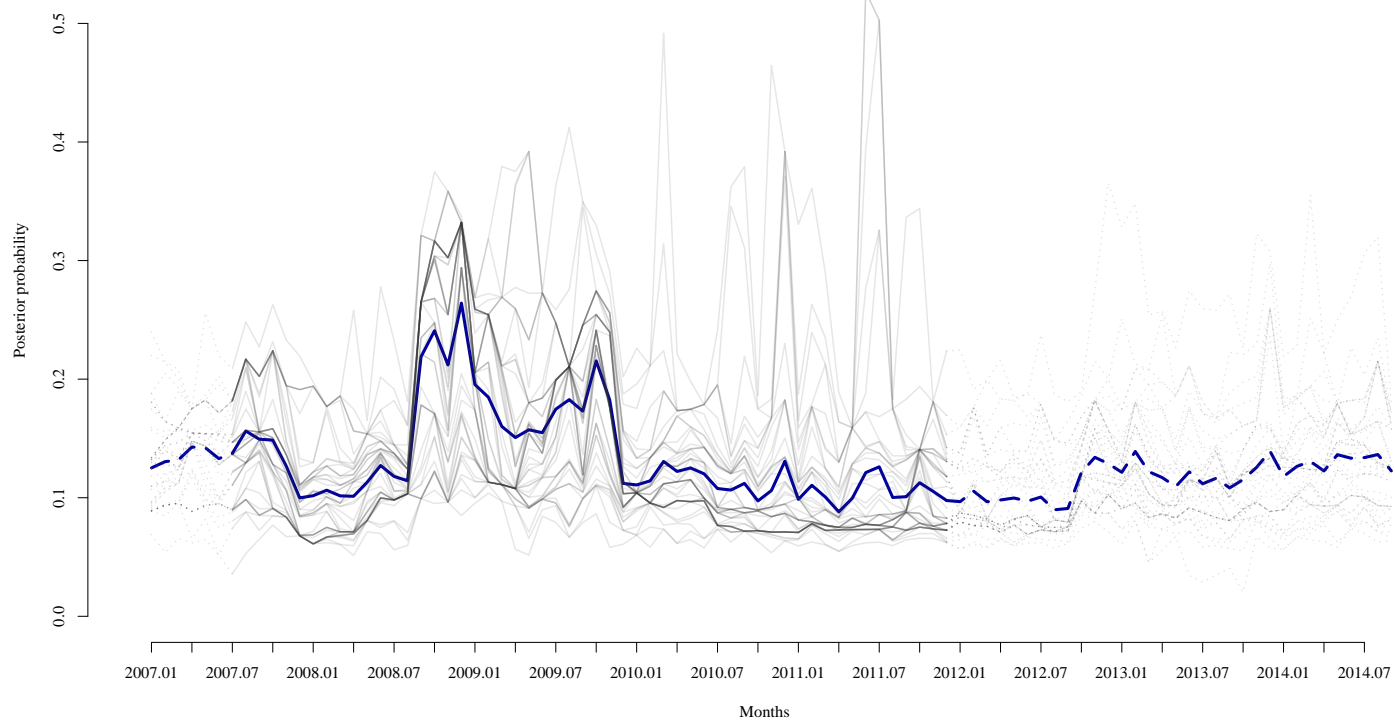
Fig. 2. Raw distress reporting. Distribution of posterior probabilities over time for sentence vectors, indicating the levels of news reporting relating to bank stress. The blue line indicates mean, faded lines every 2nd percentile, and dotted lines predictions outside the event sample.

First, Fig. 2 provides an overview of the raw distress reporting in Europe over the recent years, in terms of distributions of posterior probabilities of the sentence vectors, illustrated through their percentiles. The time span July 2007 to June 2012 is covered by the event data, and the rest is produced by applying the trained model. This distribution communicates the dynamics of the stress situation in Europe, while the mean (index I'' of Eq. 7) summarizes the general trends.

The index shows a sharp double peak starting September 2008, which coincides with the outbreak of the financial crisis. Prior to the most significant peaks, one can also observe elevated values between August and October 2007, pointing to early discussion on the significance of subprime activities overall and liquidity in European banks. The outbreak of the financial crisis in 2008 is followed by over a year of relatively high stress, where a substantial part of the cross section is elevated. A second significant and similar peak of the stress index is reached in October 2009. At the end of 2010 and 2011, one can observe notable jumps in the most extreme percentiles, whereas the rest of the cross section remains largely unaffected.

At a general level, we observe that the peak in September 2008 relates to overall distress in financial markets due to the collapse of Lehman Brothers in mid-September. However, the fact that values at the top of the distribution appear rather unstable from month to month reflects that different banks are being mentioned over time and usually not persistently across months in distress contexts. By observing increases and peaks in the index of an individual bank or banks in a country, we can identify specific events of possible relevance to distress.

4.5 Country-level stress, descriptions and interpretation

From the general stress index for Europe, this section moves to a more granular perspective on stress, closer to the level of the events being modeled. We measure stress-related discourse for countries for a more targeted stress measure, which also allows for more economic interpretation of developments, as we study the top-ranking excerpts at key points. Thus, we now aggregate posterior probabilities over time for sentence vectors, indicating the levels of news reporting relating to bank stress, but selectively at a country level (according to Eq. 6). Fig. 3 shows the developments in stress-related discussion for Belgium and Ireland and Fig. 4 for Germany and UK. The figures illustrate stress levels as time series, as well as they annotate peaks of distress levels with top-ranked excerpts. In the appendix, we include plots in Figs. 6 and 7 for the other countries whose banks we model.

In Fig. 3, the stress levels for Belgium peak in September 2008. Looking at top-ranked excerpts, September 27 is coupled with a range of rumours in media, yet no official release or actions to mitigate the weakened position of particularly Fortis Bank. Then, the next days we see a bailout of Fortis being discussed as the Belgian, Dutch and Luxembourg governments rescued Fortis. Likewise, the lower chart for Ireland in Fig. 3 shows increased concerns over Bank of Ireland and
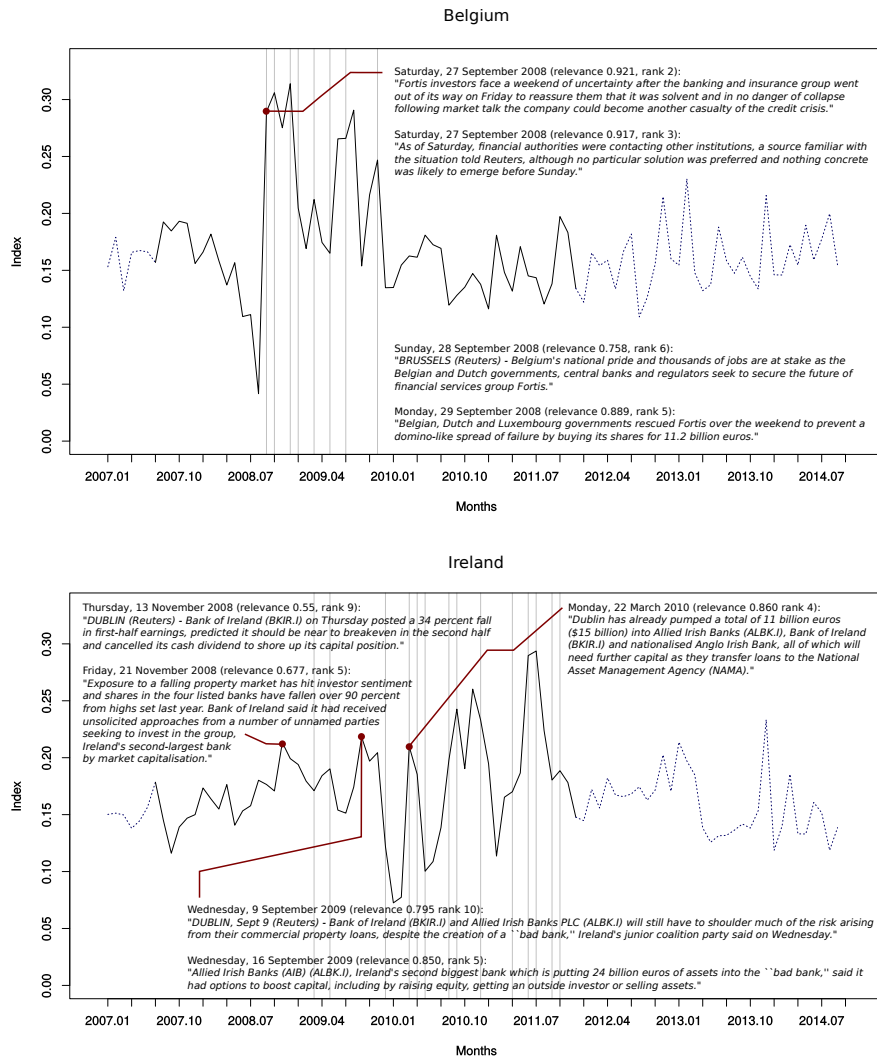
Belgium

Saturday, 27 September 2008 (relevance 0.921, rank 2):
"Fortis investors face a weekend of uncertainty after the banking and insurance group went out of its way on Friday to reassure them that it was solvent and in no danger of collapse following market talk the company could become another casualty of the credit crisis."

Saturday, 27 September 2008 (relevance 0.917, rank 3):
"As of Saturday, financial authorities were contacting other institutions, a source familiar with the situation told Reuters, although no particular solution was preferred and nothing concrete was likely to emerge before Sunday."

Sunday, 28 September 2008 (relevance 0.758, rank 6):
"BRUSSELS (Reuters) - Belgium's national pride and thousands of jobs are at stake as the Belgian and Dutch governments, central banks and regulators seek to secure the future of financial services group Fortis."

Monday, 29 September 2008 (relevance 0.889, rank 5):
"Belgian, Dutch and Luxembourg governments rescued Fortis over the weekend to prevent a domino-like spread of failure by buying its shares for 11.2 billion euros."

Ireland

Thursday, 13 November 2008 (relevance 0.55, rank 9):
"DUBLIN (Reuters) - Bank of Ireland (BKIR.I) on Thursday posted a 34 percent fall in first-half earnings, predicted it should be near to breakeven in the second half and cancelled its cash dividend to shore up its capital position."

Friday, 21 November 2008 (relevance 0.677, rank 5):
"Exposure to a falling property market has hit investor sentiment and shares in the four listed banks have fallen over 90 percent from highs set last year. Bank of Ireland said it had received unsolicited approaches from a number of unnamed parties seeking to invest in the group, Ireland's second-largest bank by market capitalisation."

Monday, 22 March 2010 (relevance 0.860 rank 4):
"Dublin has already pumped a total of 11 billion euros ($15 billion) into Allied Irish Banks (ALBK.I), Bank of Ireland (BKIR.I) and nationalised Anglo Irish Bank, all of which will need further capital as they transfer loans to the National Asset Management Agency (NAMA)."

Wednesday, 9 September 2009 (relevance 0.795 rank 10):
"DUBLIN, Sept 9 (Reuters) - Bank of Ireland (BKIR.I) and Allied Irish Banks PLC (ALBK.I) will still have to shoulder much of the risk arising from their commercial property loans, despite the creation of a ``bad bank,'' Ireland's junior coalition party said on Wednesday."

Wednesday, 16 September 2009 (relevance 0.850, rank 5):
"Allied Irish Banks (AIB) (ALBK.I), Ireland's second biggest bank which is putting 24 billion euros of assets into the ``bad bank,'' said it had options to boost capital, including by raising equity, getting an outside investor or selling assets."

Fig. 3. Distress index for Belgium and Ireland, with key periods marked and informative excerpts selected from the top-10 of each period and country. Vertical lines indicate distress events and dotted lines out-of-sample predictions. Quotes are from Reuters at given dates.
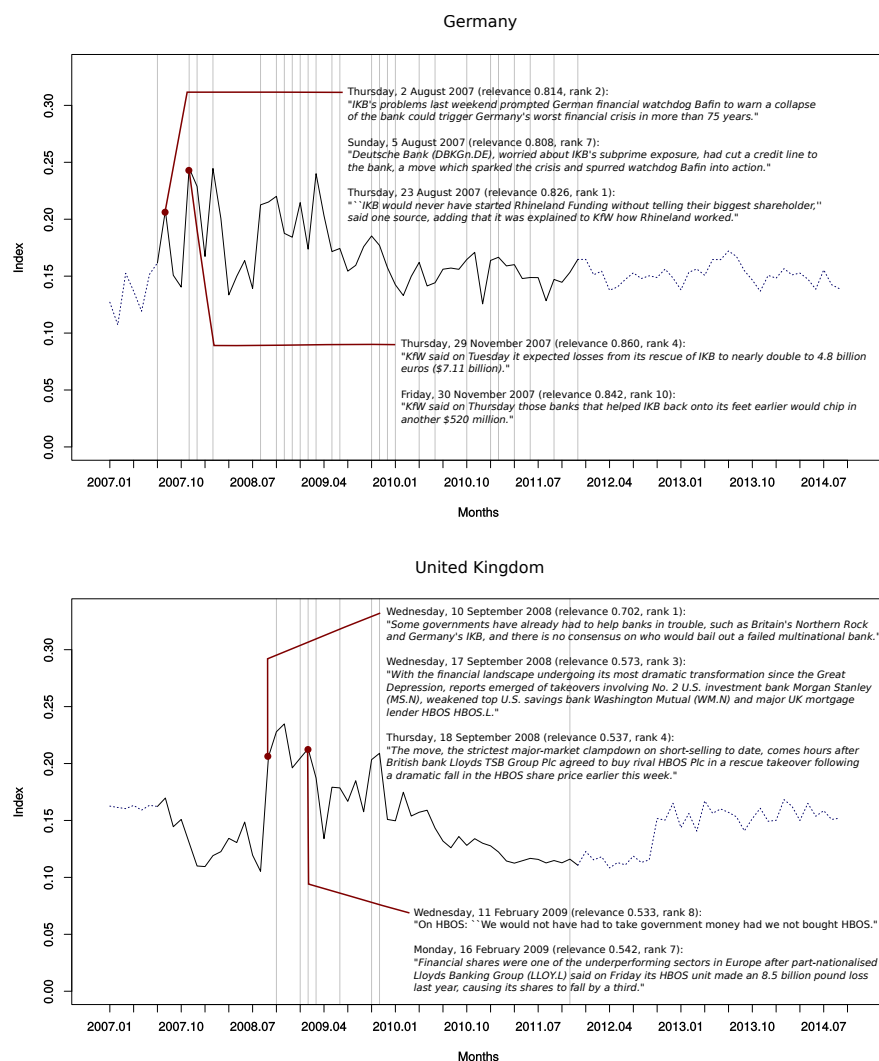
Germany

Thursday, 2 August 2007 (relevance 0.814, rank 2):
"*IKB's problems last weekend prompted German financial watchdog Bafin to warn a collapse of the bank could trigger Germany's worst financial crisis in more than 75 years.*"

Sunday, 5 August 2007 (relevance 0.808, rank 7):
"*Deutsche Bank (DBKGn.DE), worried about IKB's subprime exposure, had cut a credit line to the bank, a move which sparked the crisis and spurred watchdog Bafin into action.*"

Thursday, 23 August 2007 (relevance 0.826, rank 1):
"``IKB would never have started Rhineland Funding without telling their biggest shareholder,'' *said one source, adding that it was explained to KfW how Rhineland worked.*"

Thursday, 29 November 2007 (relevance 0.860, rank 4):
"*KfW said on Tuesday it expected losses from its rescue of IKB to nearly double to 4.8 billion euros ($7.11 billion).*"

Friday, 30 November 2007 (relevance 0.842, rank 10):
"*KfW said on Thursday those banks that helped IKB back onto its feet earlier would chip in another $520 million.*"

United Kingdom

Wednesday, 10 September 2008 (relevance 0.702, rank 1):
"*Some governments have already had to help banks in trouble, such as Britain's Northern Rock and Germany's IKB, and there is no consensus on who would bail out a failed multinational bank.*"

Wednesday, 17 September 2008 (relevance 0.573, rank 3):
"*With the financial landscape undergoing its most dramatic transformation since the Great Depression, reports emerged of takeovers involving No. 2 U.S. investment bank Morgan Stanley (MS.N), weakened top U.S. savings bank Washington Mutual (WM.N) and major UK mortgage lender HBOS HBOS.L.*"

Thursday, 18 September 2008 (relevance 0.537, rank 4):
"*The move, the strictest major-market clampdown on short-selling to date, comes hours after British bank Lloyds TSB Group Plc agreed to buy rival HBOS Plc in a rescue takeover following a dramatic fall in the HBOS share price earlier this week.*"

Wednesday, 11 February 2009 (relevance 0.533, rank 8):
"On HBOS: ``We would not have had to take government money had we not bought HBOS.*"

Monday, 16 February 2009 (relevance 0.542, rank 7):
"*Financial shares were one of the underperforming sectors in Europe after part-nationalised Lloyds Banking Group (LLOY.L) said on Friday its HBOS unit made an 8.5 billion pound loss last year, causing its shares to fall by a third.*"

Fig. 4. Distress index for Germany and the United Kingdom, with key periods marked and informative excerpts selected from the top-10 of each period and country. Vertical lines indicate distress events and dotted lines out-of-sample predictions. Quotes are from Reuters at given dates.

other large Irish banks in November 2008, as both their earnings and shares were significantly falling. After a range of actions by the state, distress levels were still peaking in September 2009, which is particularly related to the amounts that Allied Irish Banks was putting into the Irish "bad bank". Still, in March 2010 three large Irish banks were still transferring large loans to the National Asset Management Agency (NAMA). Thereafter the most acute stress decreased and has since been at lower levels, although remaining somewhat volatile.

Fig. 4 provides similar stress time series and top-ranked excerpts, but for Germany and the UK. Germany can be seen to signal already in August 2007, when IKB's problems were highlighted to potentially lead to "Germany's worst financial crisis in more than 75 years". Three days after this news, Deutsche Bank cut a credit line to IKB, as they were worried about IKB's subprime exposures, which further triggered distress in the German banking sector. One reason to the failure of IKB related to an offshore portfolio that was kept off IKB's balance sheet by Rhineland Funding, which is said to have been explained to the largest shareholder KfW. The same large shareholder is then a few months later involved in helping IKB back on its feet with a hefty 4.8 billion euros, as well as additional smaller support afterwards. For the UK, stress increased in September 2008, relating not only to previous aid to the UK-based Northern Rock but also to Germany's IKB. Here, we see an example of cross-border, systemic effects of bank distress. Only a few days later in conjunction with a strict clampdown on short-selling, UK-based bank Lloyds Group bought rival HBOS in a rescue takeover. Ironically, a few months later in February 2009 Lloyds in partly nationalized as its HBOS unit made an 8.5 billion pounds loss the year before.

4.6   The case of Fortis and IKB Bank

This section takes a final step towards more granular output by providing a stress measure for individual banks (according to Eq. 5). As with the country-level aggregates, we can aggregate posterior probabilities for sentence vectors selectively by bank. This output could be derived for each of the 101 banks, although here we focus on the stress reporting for two banks, namely Fortis and IKB Bank.

One of the early failures among European financial institutions occurred to the Benelux-based Fortis. As was also highlighted in the above described top excerpts for Belgium, Fortis and the rescue procedure was at the core of the discussion as the crisis erupted. We focus on the evolution of the distress index for Fortis, as is shown in Fig. 5. To start with, we can observe that elevated values for the stress index coincide with distress events.

By the first event in September 2008, the index rises to 0.30, which marks the start of a prolonged period of elevated stress. The top-ranked excerpts relate to a range of different issues, such as worries about lacking confidence in the markets and the systemic nature of the unfolding crisis:

"Jean-Claude Juncker, also the prime minister of Luxembourg, was asked whether the part nationalisation of Dutch-Belgian bank Fortis FOR. -BR and a new injection of liquidity into money markets by the European Central Bank would restore market confidence. "I can only hope that confidence will come back – financial markets should not forget to take a close look at the health of fundamental data of several banks – and that this casino game, that's going on independently from the good fundamentals, stops," he told reporters on the sidelines of a meeting in parliament. Belgian, Dutch and Luxembourg governments rescued Fortis over the weekend to prevent a domino-like spread of failure by buying its shares for 11.2 billion euros."
(Reuters 2008-09-29, relevance 0.963, rank 1)

"Investors also worried if a proposed U.S. rescue would stem the contagion that pushed the British government to takeover troubled mortgage lender Bradford & Bingley BB.L and three European governments to partially nationalize banking and insurance group Fortis FOR.BRFOR.AS."
(Reuters 2008-09-29, relevance 0.923, rank 6)
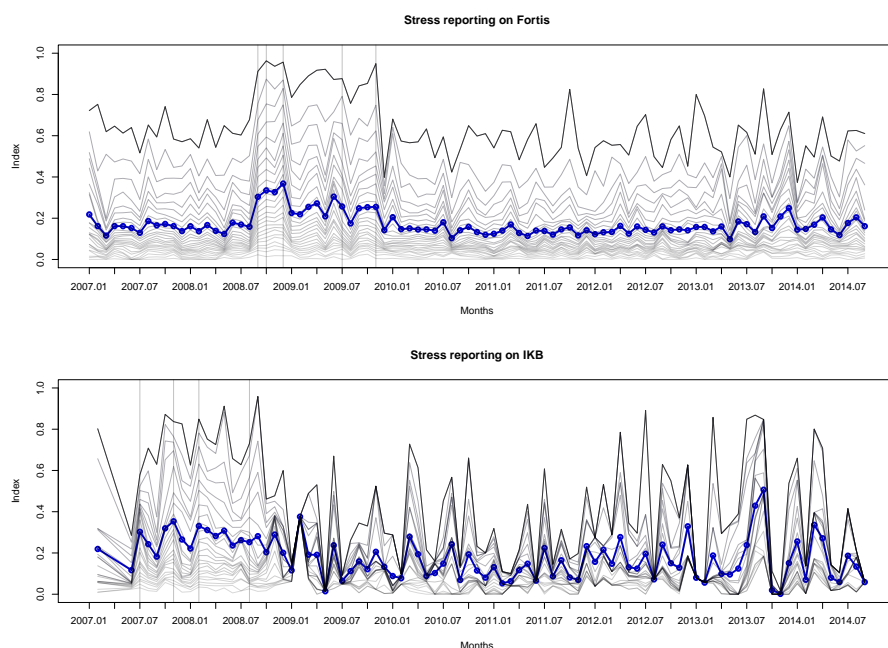
Fig. 5. Indices (blue) for banks Fortis and IKB indicating the levels of bank stress-related reporting, with faded lines showing every 4th percentile up to the 98th. Vertical lines indicate recorded events.

In October 2008, the top excerpts discuss the continuing developments such as the Benelux governments "carving up" Fortis to sell to private entities, including French BNP Paribas buying control of the arms in Belgium and Luxembourg. Further excerpts highlight the cross-border aspect of the interventions, and the issues it entails:

> *"The Fortis deal is the biggest cross-border rescue since the full force of the credit crisis swept across the Atlantic into Europe last month, upending banks and rattling saver confidence."*
> (Reuters 2008-10-06, relevance 0.945, rank 7)

> *"Dutch Finance Minister Wouter Bos fanned Belgian resentment by telling journalists: 'Many of the problems were hidden in the Belgian part of the Fortis group.'"*
> (Reuters 2008-10-05, relevance 0.945, rank 8)

This repeats the message of the already cited news for the UK in September 2008, that "there is no consensus on who would bail out a failed multinational bank", highlighting how the use of text descriptions can provide deeper insight into the multifaceted developments underlying a model signal.

Without a detailed analysis of the discussion around the IKB Bank, we can again conclude from Fig. 5 that the stress index takes high values during the realized events. Generally, the top-ranked discussion herein correlates to a large extent with the early top-ranked discussion for Germany, as was above exemplified. The discussion around the distress events relates to early indications of stress, ties to other German banks and government actions taken during and after the stress episodes. After a period of elevated stress during 2007–2008, the figure illustrates that stress is still fairly volatile and that the most extreme percentiles still take large values. This may relate to the fact that discussion keeps relating to the 2007–2008 distress events, in that the solution to the stress events was an acquisition by a investment company. The private equity firm Lone Star acquired IKB Bank in 2008 with the aim of restructuring and selling the bank, and accordingly any rumours still link it to the original stress discussion during the global financial crisis. Such references to past major stress events may however also be an indication of current concerns about financial stress, thus worth signaling in order to allow further investigation.

5   Conclusions

We have presented a deep-learning-based approach that combines two types of data, news text and basic event information, with the aim of linking the two to describe observed and predicted events. The approach entails unsupervised learning on text in order to model its language and provide semantic vector representations that are used as features for predictive modeling of events. The neural-network-based method that we put forward is able to work with a very small set of events, matched with text through a heuristic, in order to discern

what type of language and passages in the text are actually relevant to the modeled event type and phenomenon. The semantic modeling utilizes large amounts of text data to infer abstractions that counter the high variability and sparsity of language, thus supporting prediction of infrequent events.

The semantic-predictive model can produce indices that indicate the level of relevant discussion over time, overall or related to specific entities or groups thereof. The indices can highlight interesting patterns and offer guidance in the search for relevant events, whereas the model very directly provides means to rank and retrieve pieces of text from news articles in order to describe the quantitative signal.

We demonstrate the usefulness of the method and the possibilities of the approach in general within the study of financial risk, by modeling bank distress events. The indices reflect the level of current reporting related to bank stress over time at multiple levels: for Europe in general, for individual countries and for specific banks. Guided by the indices, users may focus their search and retrieve the relevant reporting of the time, in order to understand the developments regarding, in this case, government interventions and rescues. Our quantitative evaluation of the stress index shows good results and provides an important quality assurance of the descriptions.

The method and our analysis exemplify how text may offer an important complementary source of information for financial and systemic risk analytics, which is readily available, current and rich in descriptive detail. In contrast to traditional information sources, text data offers a possible route to circumvent the issues of privileged access, lagging publication and low granularity, but most importantly does it very directly offer value through the explanatory power of the event-related human language descriptions accompanying the plane signal. We expect the method to be also directly applicable to describe events beyond the financial domain, relating to geopolitics and other significant topics.

We recognize that deep learning approaches are useful in particular to handle the complexities of such new types of data, while offering necessary flexibility when exploring new fields of analysis. Seeking to harness the expressiveness of text, we should continue to look to computational linguistics for support in terms of theoretical foundations and tools.

While we show that it is possible to predict relevance and retrieve informative descriptions of events, we merely scratch the surface of the vast text material in any given cross section with our current method of presentation. A challenge remains in developing methods that are able to meaningfully summarize the broader base that may include a long tail of weakly signaling, subtle expressions. Such signals may be particularly important in order to register and track developments before they materialize in severe and obvious events. Likewise, to really make use of text data as a complement rather than a replacement, traditional sources and text should be integrated in a unified modeling framework in order to achieve the best predictive performance possible, while also keeping the opportunity to explore the descriptions to that signal open.

6   Appendix

Figs. 6 and 7 provide country-level indices for the countries not included in Figs. 3 and 4, as well as the non-weighted average of all country indices. The individual banks and countries they are mapped to are listed in Table 3.
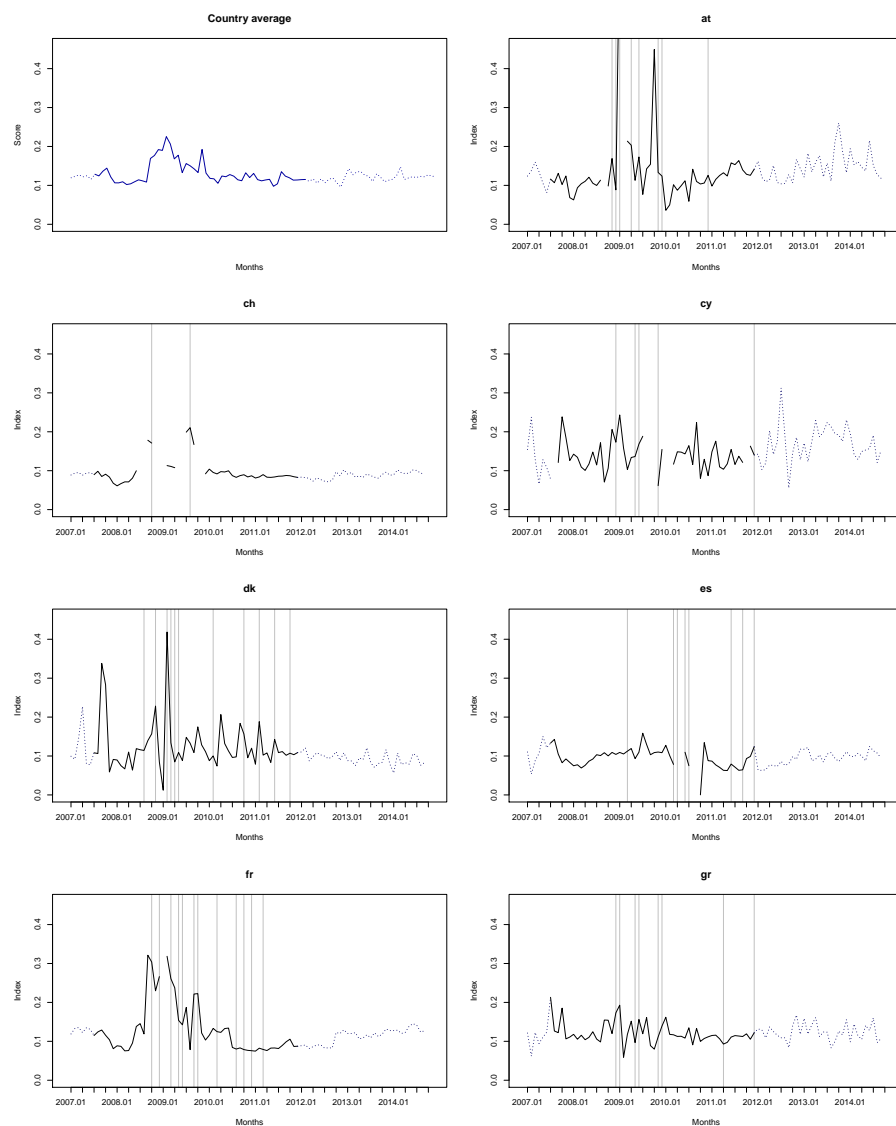
Acknowledgment

Fig. 6. Distress index for Austria, Switzerland, Cyprus, Denmark, Spain, France, Greece and average of all modeled countries. Vertical lines indicate bank-level distress events and dotted lines out-of-sample predictions.
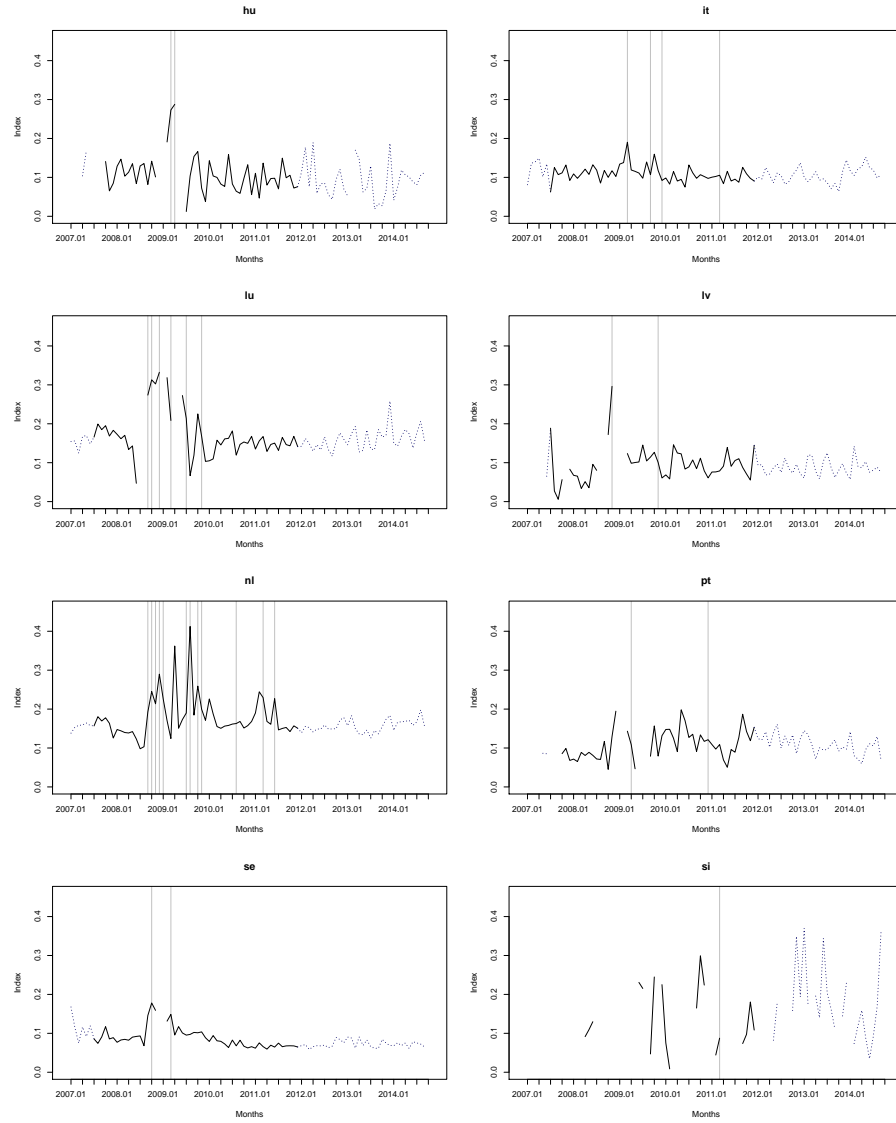
Fig. 7. Distress index for Hungary, Italy, Luxembourg, Latvia, Netherlands, Portugal, Sweden and Slovenia. Vertical lines indicate bank-level distress events and dotted lines out-of-sample predictions.

Bank	Country	Bank	Country
ABN Amro	NL	Fionia (Nova Bank)	DK
ATE Bank	GR	First Business Bank	GR
Aareal Bank	DE	Fjordbank Mors	DK
Aegon	NL	Fortis Bank	LU, NL, BE
Agricultural Bank of Greece	GR	HBOS	UK
Allied Irish Banks	IE	HSH Nordbank	DE
Alpha Bank	GR	Hellenic	GR
Amagerbanken	DK	Hypo Alpe Adria Group	AT
Anglo Irish Bank	IE	Hypo Real Estate	DE
Attica Bank	GR	Hypo Tirol Bank	AT
BBK	ES	IKB	DE
BNP Paribas	FR	ING	NL
BPCE	FR	Irish Life & Permanent	IE
BPP	PT	Irish Nationwide Building Society	IE
Banca Civica	ES	KBC	BE
Banca Popolare	IT	Kommunalkredit	AT
Banca Popolare di Milano	IT	LBBW	DE
Banco Mare Nostrum	ES	Lloyds TSB	UK
Banco Popolare	IT	Lokken	DK
Banco de Valencia	ES	Magyar Fejlesztesi Bank Zrt	HU
Bank of Cyprus	CY	Marfin Popular Bank	CY
Bank of Ireland	IE	Max Bank	DK
Bankia	ES	Monte dei Paschi di Siena	IT
Banque Populaire	FR	Mortgage and Bank of Latvia	LV
Bawag	AT	National Bank of Greece	GR
BayernLB	DE	NordLB	DE
CAM	ES	Nordea	SE
Caisse d'Epargne	FR	Northern Rock	UK
Caixa General de Depositos	PT	Nova Ljubljanska banka	SI
Caja Castilla-La Mancha	ES	Novacaixagalicia	ES
Caja Espana	ES	OTP Bank Nyrt	HU
Carnegie Investment Bank	SE	Panellinia Bank	GR
Catalunyacaica	ES	Pantebrevsselskabet	DK
Commerzbank	DE	Parex	LV
Cooperative Central Bank	CY	Piraeus Bank	GR, CY
Credit Agricole	FR	Proton Bank	GR
Credit Mutuel	FR	RBS	UK
Credito Valtellinese	IT	RZB Group	AT
Cyprus Development Bank	CY	Roskilde Bank	DK
Cyprus Popular	CY	SNS Reaal	NL
Danske Bank	DK	SachsenLB	DE
Dexia	BE, FR, LU	Societe Generale	FR
Dunfermline	UK	Swedbank	SE
EBH	DK	T-Bank	GR
EBS Building Society	IE	UBS	CH
EFG Eurobank	GR	UNNIM	ES
Eik Bank	DK	USB Bank	CY
Erste Bank	AT	VBAG	AT
Ethias	BE	Vestjysk	DK
FHB Jelzalogbank Nyrt	HU	WestLB	DE
Finansieringsselskabet	DK		

Table 3. Target banks and their countries.

References

1. J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

2. B. Back, T. Laitinen, and K. Sere. Neural networks and genetic algorithms for bankruptcy predictions. *Expert Systems with Applications*, 11(4):407 – 413, 1996.

3. M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247, 2014.

4. Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

5. F. Betz, S. Oprică, T.A. Peltonen, and P. Sarlin. Predicting distress in european banks. *Journal of Banking & Finance*, 45:225–241, 2014.

6. 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.

7. J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP'09 Shared Task on Event Extraction*, pages 10–18, 2009.

8. J. Borsje, F. Hogenboom, and F. Frasincar. Semi-automatic financial events discovery based on lexico-semantic patterns. *International Journal of Web Engineering and Technology*, 6(2):115–140, 2010.

9. P. Capet, T. Delavallade, T. Nakamura, A. Sandor, C. Tarsitano, and S. Voyatzi. A risk assessment system with automatic extraction of event types. In *International Conference on Intelligent Information Processing*, pages 220–229. Springer, 2008.

10. P. Cerchiello, P. Giudici, and G. Nicola. Big data models of bank risk contagion. DEM Working Paper Series, 117 (02-16), 2016.

11. R.A. Cole and J.W. Gunther. Predicting bank failures: A comparison of on- and off-site monitoring systems. *Journal of Financial Services Research*, 13:103–117, 1998.

12. A.I. Dimitras, R. Slowinski, R. Susmaga, and C. Zopounidis. Business failure prediction using rough sets. *European Journal of Operational Research*, 114(2):263 – 280, 1999.

13. R. Gropp, J. Vesala, and G. Vulpes. Equity and bond market signals as leading indicators of bank fragility. *Journal of Money, Credit and Banking*, 38(2):399–428, 2006.

14. Z.S. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

15. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

16. F. Hogenboom, M. de Winter, F. Frasincar, and U. Kaymak. A news event-driven approach for the historical value at risk method. *Expert Systems with Applications*, 42(10):4667–4675, 2015.

17. F. Hogenboom, F. Frasincar, U. Kaymak, F. de Jong, and E. Caron. A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85:12 – 22, 2016.

18. J. Hokkanen, T. Jacobson, C. Skingsley, and M. Tibblin. The Riksbank's future information supply in light of Big Data. In *Economic Commentaries*, volume 17. Sveriges Riksbank, 2015.

19. O. Irsoy and C. Cardie. Deep recursive neural networks for compositionality in language. In *Advances in Neural Information Processing Systems*, pages 2096–2104, 2014.

20. Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

21. A. Lischinsky. In times of crisis: a corpus approach to the construction of the global financial crisis in annual reports. *Critical Discourse Studies*, 8(3):153–168, 2011.

22. 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):782–796, 2014.

23. K. Männasoo and D.G. Mayes. Explaining bank distress in Eastern European transition economies. *Journal of Banking & Finance*, 33:244–253, 2009.

24. 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.

25. A. Milne. Distance to default and the financial crisis. *Journal of Financial Stability*, 12:26–36, 2014.

26. J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

27. Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

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. T. Peltonen, A. Piloui, and P. Sarlin. Network linkages to predict bank distress. ECB Working Paper, No. 1828, 2015.

30. S. Rönnqvist and P. Sarlin. Bank networks from text: Interrelations, centrality and determinants. *Quantitative Finance*, 15(10), 2015.

31. S. Rönnqvist and P. Sarlin. Detect & describe: Deep learning of bank stress in the news. In *Proceedings of IEEE Symposium Series on Computational Intelligence*, pages 890–897, 2015.

32. D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

33. P. Sarlin. On policymakers' loss functions and the evaluation of early warning systems. *Economics Letters*, 119(1):1–7, 2013.

34. J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

35. H. Schütze. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, Supercomputing '92, pages 787–796, Los Alamitos, CA, USA, 1992. IEEE Computer Society Press.

36. H. Schütze and J. Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.

37. R. Socher and C. Manning. Deep learning for natural language processing (without magic). Keynote at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013). http://nlp.stanford.edu/courses/NAACL2013/.

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

39. K.S. Tai, R. Socher, and C.D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Association for Computational Linguistics (ACL)*, 2015.

40. H. Tanev, J. Piskorski, and M. Atkinson. *Real-Time News Event Extraction for Global Crisis Monitoring*, pages 207–218. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

41. C.J. Van Rijsbergen. *Information Retrieval*. Butterworth, 2nd ed., 1979.

42. W.Y. Wang and Z. Hua. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014.

43. Y. Yang, J.G. Carbonell, R.D. Brown, T. Pierce, B.T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.