Recognizing Weak Signals in News Corpora

Daniela Gîfu

[1]„Alexandru Ioan Cuza” University of Iași, Faculty of Computer Science,
daniela.gifu@info.uaic.ro

Abstract. This paper presents a suit of experiments carried out with different machine learning systems developed to learn how to find weak signals in news corpora. The task is particularly challenging as the weak signals are not marked at word or sentence level, but rather at document/paragraph level and while there is no explicit definition that strictly applies to them, people are very good at recognizing them. Purposely lacking strict annotation guidelines, a large news corpus was annotated via tacit knowledge and then we used a supervised learning technique to reproduce the weak signal label. We report a large improvement in weak signals recognition using deep learning over other approaches, useful for investors, entrepreneurs, economists, and normal users, to give them a clue on how to invest.

Keywords: deep learning, weak signals, news corpora, statistics

1 Introduction

How would be it like going back in time and reading scientific predictions from 10 years ago? Some of them would sound very strange, which supports once again the Old Danish proverb that “making predictions is hard, especially about the future”. However, sometimes, some predictions may prove quite accurate [Popescu & Strapparava, 2013; Gifu, 2015]. A subset of these are the result of corroborating small evidence existing in separate places for which the necessity to put them together was found by some visionary capable minds. Concisely, this paper is about building NLP (*Natural Language Processing*) systems to improve the probability for this type of predictions to be true.

The need for detecting weak signals, defined as imprecise and early indicators of impending important trends or events, has been under focus in the last years. It becomes critical to realize what the next trends in economy will be. In science, making predictions almost equates to having a bright idea on how apparently disparate small achievements may be brought together to make something that is very promising.

We present a methodology able to learn to discern between what may be *vs.* what is not a weak signal, and to retrieve and corroborate document that seems to amplify it.

Initially we approach this problem as a supervised task. We compiled a relative large corpus of news from scientific journals, and we relied on a group of annotators to mark paragraphs in news from scientific journals, which may contain weak signals according to the annotators’ judgment. Then we run a series of learning algorithms to see the extent to which we could automatically identify what paragraph may contain weak signals. As it turned out that we can do it with a relatively good accuracy, we were wondering whether we could learn recognizing weak signals by reducing dependency on annotated data. To reach this aim we used Google Ngrams Viewer to analyze the time distribution over certain topics we identified automatically from the corpus. We selected some of the topics described by an acceding ratio following a technology similar to the one presented in Popescu & Strapparava [2014] or Amarandei *et al*. [2017]. Our assumption is that the description of these topics at some point in time before they had a large impact was made by using weak signals. We automatically compiled a corpus from the time just before these topics became clearly important and we constructed word embedding for what appeared to be the common vocabulary in describing these topics. We trained a deep learning algorithm based on gradient descent and we implemented a LSTM (*Long Short-Term Memory*) neural network. Next, we tested these systems on a set of documents for which we knew whether they contained weak signals or not, thanks to the previous supervised experiment. The results showed that we could learn to make predictions in an unsupervised manner, based on unknown weak signals, obtaining an accuracy, which is close to the supervised one. This is the fundamental result reported in this paper.

2 Related Work

The literature on weak signals is not very large, as this field is about to emerge. A groundbreaking paper [Brynielsson, *et al.*, 2012; Cohen *et al*., 2014] was looking mainly to weak signals for detecting deviational behavior to efficiently provide preemptive counter measures. The probabilistic model presented here is similar to the one used in language modelling (an estimation of posterior probability of certain class via chain formula).

In [Wang *et al.*, 2012] an automatic detection of crime using tweets is presented. They use LDA (*Latent Dirichlet Allocation*) to predict classes of similar words for topics related to violence.

While we can gain a valuable insight from these papers, their scope is limited because there is a direct connection between the overt information existing in text and speaker's intention.

Another emerging field, diachronicity (i.e. the evolution of certain topics in mass media over time) is linked to detection of weak signals. We found useful the statistical tests presented for epoch detection in [Mihalcea & Nastase, 2012], or temporal dynamics in [Wang & McCallum, 2006; Wang *et al.*, 2008; Gerrish and Blei, 2010].

In [Hastie *et al.*, 2008; Mustafa, 2012] we found very useful insights from dealing with discriminative analysis and SVM (*Support Vector Machine*) respectively. In order to improve our results we had to understand how we could restrict further the objective function. In addition [Rocktaeschel, 2016], the principle of an attentive neural network is presented.

3 Materials & Methods

We describe the data set and different machine learning systems used in to predict how to find weak signals in news.

3.1 Data Set

We started by compiling a corpus of scientific papers and articles (42,916) from on-line freely available repositories published between 1960 and 2017.

Our assumption is that weak signals represent a form of tacit knowledge. As such, it may be counterproductive to define a formal guidelines aiming to identify the weak signal. Rather, we let the annotator the liberty to mark a whole document as containing weak signals or not. In a second round of annotations, we wanted to restrict the scope to paragraph rather than whole document. Most of the annotated paragraphs have between 100 to 250 words.

Therefore, we obtained two annotated corpora, which for convenience we refer to short and long respectively. The long corpora, LC, refer to full documents as training/test corpora. The short corpora, SC, refer to paragraphs.

There is not a perfect overlap between these two corpora; approximatively 15% of paragraphs come from different documents that the ones considered on LC corpora.

The annotation is binary, YES or NO signaling the existence of weak signals or not respectively. In case of SC, all paragraphs, not explicitly classified as "YES" from the analyzed documents, were be considered "NO". However, we double-checked the SC "NO" for some of these paragraphs in order to make sure that there are as little as possible misclassifications. We have the following distribution:

Table 1. Size of hand annotated corpora.

	YES	**NO**
LC	4,100	14,020
SC	3,700	14,500

We wanted to have a similar ratio of weak *vs.* non weak in both corpora for easing a fair comparison of the performances for these two corpora. The team consists of 18 undergrads volunteers. On a given 300 documents the annotators were encouraged to discuss their doubts and to defend their position in case of disagreement.

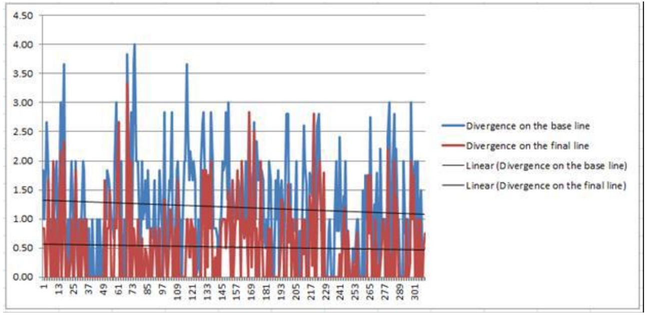


Fig. 1. Towards reaching a stable shared tacit knowledge

In Fig.1, we plot the evolution of the average number of documents on which there was a strong disagreement, for samples of 10 documents out of the chosen 300. The average disagreement lowered from 1.4 to 1.1 and the divergence decreased from .55 to .38.

It seems that 1.1 is a hard threshold for this task. When we repeated the experiment after we had 1,200 of documents annotated as carriers of weak signals, the average of disagreement for samples of ten documents, was still 1.1. For making a decision decreased for time between these two experiments we measure the average time (Fig. 2).

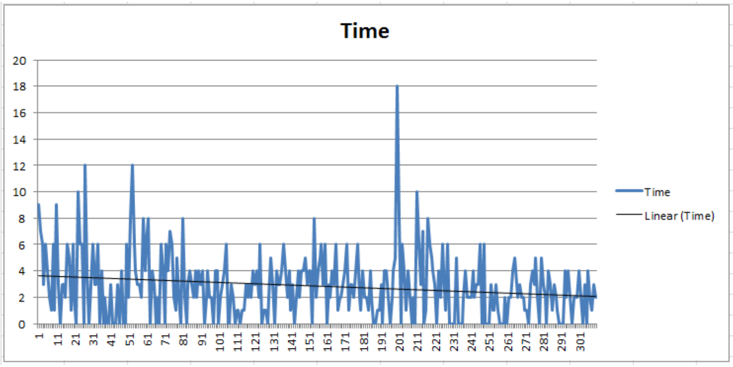
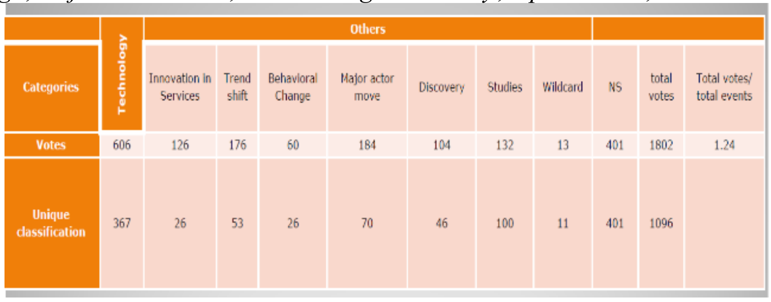


Fig. 2. Average time for making a decision

These results suggest that this task, in spite of being one driven by tacit knowledge, is learnable by algorithmic probabilistic hypothesis space search. The annotators developed patterns – they seem to filter out a lot of the content, otherwise the time to reach a decision would not have decreased that dramatically, and there is gray zone where experience does not help. This behavior tends to help an automatic classifier, as it does not have to be very precise in order to obtain a human similar performance.

After a preliminary round of trial annotation of several hundred of documents, we decided to create a taxonomy that sprung naturally from this experiment based on the a several components: *technology*; *innovation in services*; trend shift; *behavioural change*; *major actor move*; *breakthrough discovery*; *top research*; *wild card*.



		Others									
Categories	Technology	Innovation in Services	Trend shift	Behavioral Change	Major actor move	Discovery	Studies	Wildcard	NS	total votes	Total votes/ total events
Votes	606	126	176	60	184	104	132	13	401	1802	1.24
Unique classification	367	26	53	26	70	46	100	11	401	1096	

Fig. 3. Weak signal taxonomy distribution

The intention in using these labels was to try to capture the annotator intuition on why a certain document/paragraph is considered as carrier of weak signals. This taxonomy helps us to see if there are indeed any subjective differences that may affect the learning process. It came as a surprise that the annotators did not want to use often the wild card taxonomy. The number of documents that received just one category is relatively high and quasi constant (50%). The number of documents that received more than three categories is non-significant, less than 3%.

In Fig. 4, we draw the dynamics of reaching consensus among annotators.

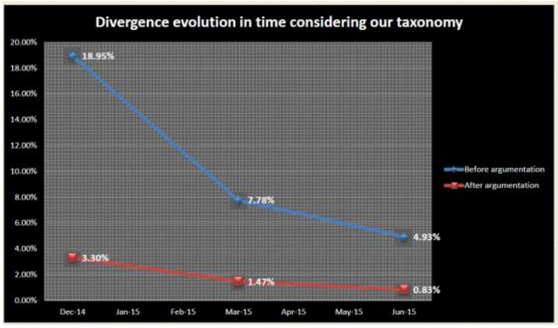


Fig. 4. Reaching consensus over taxonomies

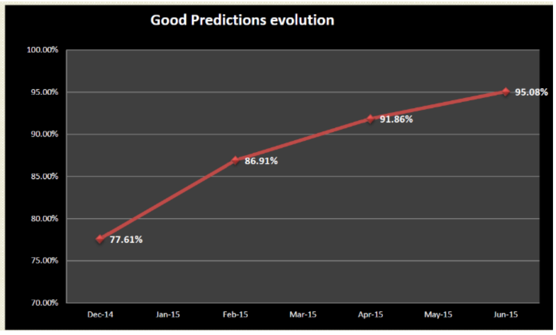


Fig. 5. Control Group Judgment

We wanted to check whether this consensus was reached due to an increasingly strong and commonly shared tacit knowledge, i.e. due to acquiring an expertise, or due to accepting a dominant view. A control group checked the validity of the agreement what we found we have found that the results strongly suggest the first alternative – acquiring an expertise.

Finally, all these experiments strongly suggest that we have a tacit knowledge about weak signals that is shared at least 80% of the time. However, there will be a 10% individual hard kernel that constitutes a potential disagreement area. Maybe this is exactly the prediction on the immediate future.

3.2 Machine Learning

We present a series of learning approaches, which we tried systematically. The experience and insight gained from previous steps guided our decision in designing the next step.

3.2.1. Baseline

In a supervised approach, finding the pieces of news containing weak signals is a binary classification task. A first approach is to use *tf-idf* weights to compute the similarity between a document and the documents in one of the two classes. This provides us with a weak baseline. However, it is an informative one. It tells how much of the weak signals are judged to be expressed via some special words or patterns. Anticipating, it turns out that this is not the case at all. This baseline has negligible accuracy, far distanced from the best results we obtained eventually. This preliminary finding confirmed that the task is not trivial at all and that many clues based on which the human judges the correct answer are not necessarily expressed by clearly defined overt phrases.

As such, we can use a couple of off the shelf approaches that will provide a set of baselines for this task. We looked at two libraries, which implement quadratic discriminative analysis, QDA (*Quadratic Discriminant Analysis*) from SCIKIT library, and SVM, linear SVM form WEKA (*Waikato Environment for Knowledge Analysis*) library, respectively. See also the equations 1 and 2.

$$P(y = k|X) = \frac{P(X|y = k)\,P(y{=}k)}{P(X)} - \frac{P(X|y = k)\,P(y{=}k)}{\sum_l P(X|y = l)\cdot P(y{=}l)} \quad (1)$$

$$\hat{\mathrm{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \pi(f(X_i) \neq Y_i \quad (2)$$

The reasons behind our choice have to do with the type of data we employ here. The fact that the *tf-idf* obtained a very low score does not immediately imply that maximizing the prior probability *P(word|weak signal)* is inefficient.

At this point, we have to understand whether the projection of the data into a bidimensional space will lead to conelike structures, that is, that the data can separated by a quadratic function. On the other hand, if the difference between the SVM and QDA is large enough this will show that QDA suffers from the masking effect.

We run both QDA and SVM in a cross-validation setting, 10 folds 1/8 ratio for train/test and 1/8 ratio for development/train.

Table 2. QDA, SVM for SC.

	Weak Signal	**No Signal**
QDAcr	0.412	0.877
SVMcr	0.663	0.913
QDAts	0.403	0.865
SVMts	0.610	0.905

That is we used a tenth of the corpus for test and development respectively. For test we used 500 weak-signals and 500 no-signals. In Table 2 we present the results for QDA and SVM for SC, and in Table 3 the results for LC for cross validation.

The *tf-idf* scored 0.18 for and 0.12 respectively. As we can see both QDA and SVM scored significantly better than that. In addition, indeed there is a non-random difference between QDA and SVM results.

Table 3. QDA, SVM for LC.

	Weak Signal	**No Signal**
QDAcr	0.38	0.901
SVMcr	0.472	0.946
QDAts	0.365	0.890
SVMts	0.455	0.930

To understand better the nature of this difference we ran a series of experiments alternating the ratio of weak signals in the training corpus. We found no significant differences from Table 2 and 3. This shows that probably we cannot improve these results by adding more training. Given that SVM is a constraint over a large boundary for ||Ein-Eout|| and that the differences from QDA are large, equation 1 follows that it is possible to search for a better model even further. That is, particularly for this task, we could find a better estimation, as the worst-case scenario seems not to characterize this corpus. Because, we cannot directly compute the number of dichotomies, and therefore, the exact VC (*Vapnik-Chervonenkis*) dimension is unknown, based on the Tables 2, 3 it is intuitively tempting to consider that the VC bound is indeed too loose for this task. That is, we can do better in estimating the posterior probability. The right question is whether we have enough data to train a more detailed classifier. We may guess that deep learning methods may be up to the task.

3.2.2. Deep Learning Approach

We describe here two experiments carried out using deep learning methods. The first method uses a simple gradient descent model with CR (*Cross Entropy Loss*) or log loss, function. The second is a LSTM neural network. The cross entropy is described by equations (3).

$$H_{y'}(y) = -\Sigma_i y_i' \log(y_i) \qquad (3)$$

LSTM

$H = [X_t\ h_{t-1}]$ $\quad$ $\hat{\imath}t = \sigma(W^iH+b_i)$

$ft = \sigma(W^fH+b_i)$ $\quad$ $ot = \sigma(W^0H+b_0)$

We organized the experiments in the following manner: we made a test corpus of 1,500 documents, 1000 non signals and 500 signals. The rest of the corpus was sent to the deep learning algorithm. Each document was represented as a vector via word embedding. We used a batch of 100 vectors, selected randomly, at one training cycle over a couple of thousands of cycles. In order to both (i) have a better assessment of the accuracy and (ii) analyze the influence of signal *vs.* non signal ratio in training we ran the algorithm three times. The number of test non-signal was 1,000 and the number of test weak-signal was 500. We varied the ration weak-signal/no-signal in one batch of 100 random input example to train. CR_n/m, represents a system with cross entropy loss function that has 2,000 cycles with a batch of 100 containing *m* random weak-signal and *n* random no-signal from training corpus. We tested for n/m in {5/5, 3/7} (rows in tables below). For example, LS_5/5 represents the system of LSTM with 100 batches made of 50 weak and 50 no signals, randomly chosen from training set.

Table 4 and 5 show the results. The columns represent the accuracy for each of the three runs for weak and no signals respectively.

Table 4. CR, LS for SC.

	1W	1N	2W	2N	3W	3N
CR_5/5	0.75	0.85	0.74	0.86	0.79	0.86
LS_5/5	0.75	0.89	0.73	0.92	0.75	0.87
CR_3/7	0.88	0.93	0.86	0.93	0.85	0.92
LS_3/7	0.86	0.94	0.86	0.95	0.88	0.95

Table 5. CR, LS for LC.

	1W	1N	2W	2N	3W	3N
CR_5/5	0.75	0.85	0.74	0.86	0.79	0.86
LS_5/5	0.75	0.89	0.73	0.92	0.75	0.87
CR_3/7	0.88	0.93	0.86	0.93	0.85	0.92
LS_3/7	0.86	0.94	0.86	0.95	0.88	0.95

As expected, the results are usually better on SC than on LC. However, the most important thing was that both approaches produces consistent results for each run, there were no big jumps/drops in accuracy at any particular experiment.

We notice a great improvement over the SVM performances, which prove that the data was enough to compute accurate Softmax estimation. It also shows that for this particular problem, the VC bound is too loose indeed.

In order to evaluate the dependency of the deep learning algorithm on the number of examples, we ran a second round of experiments considering for training only q half, two thirds, 3 quarts and 7/8 of the initial training data. It turns out that there is no statistically significant difference once we feed more than 3 quarts of the data. This means that it is unlikely that these results will be further improved by providing more data alone.

3.3 Unsupervised Learning

We present a methodology of learning weak signals in an unsupervised way. In the previous section, we obtained very good results via training examples. The annotated data, in spite of coming by a long time consuming process, may still contain errors, may still be subjectively biased by guidelines. A better way is to induce the training data and then fed it to the deep learning algorithm. Concisely, we are going to consider a set of topics that were introduced by weak signals before they become very important topics that everybody talks about.

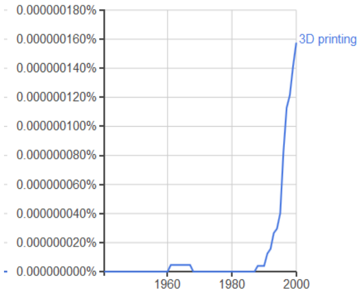


Fig. 6. 3Dprinting N-gram distribution

For example, self-driving cars, 3D printing, Higgs boson, etc. We know this set of topics as a side effect of the annotation process explained in Section 3. We use the Google N-gram to observe the diachronic evolution of a particular topic. As expected, these topics have a boom, which has a very particular steeply ascending plot starting after a certain year.

This booming trend can be captured via statistical tests. It is very easy to detect the booming period due to its clearly defined shape, but what we are actually after is the pre boom period. We take the Welch and ratio tests and we detect the pre boom statistically correlated period. In this period both the Welch and ratio test, indicate that something is going on, but there is no boom yet. That is, a non-random variation into the distribution of previous years is caught, but there is no confirmed boom yet. We assume that this is the period when that particular topic was rather described using weak signals. In a few years, after the boom, the description is far from "weak signal", it is a confirmed trend. See Fig. 7, where we marked the pre boom period we automatically found for *Higgs boson.*

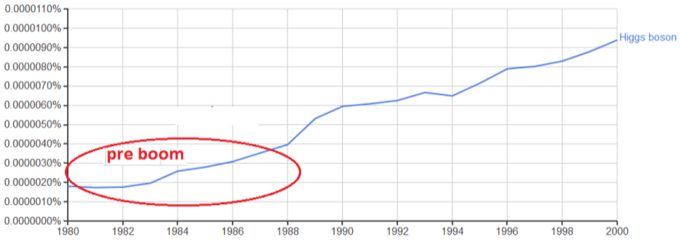


Fig. 7. Pre boom period for Higgs boson

The pre boom period is characterized by a positive ratio test followed by a period of relatively stability, thus negative Welch test.

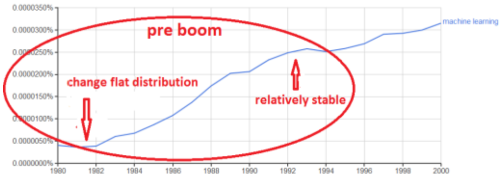


Fig. 8. Machine learning pre boom

A human interpretation of this behavior is that there are initial excitements about a certain achievement, which make people talk more often that before about it, so the old flat distribution changes. Then there is a period of relatively silence, and then there is the boom. Interestingly also machine learning was found as important topic, with a clear definite pre boom and boom period

For the four topics mentioned above, we plot in Table 7 the pre boom periods we found.

Table 7. Pre boom period.

Topics	Pre boom period
machine learning	1978-1990
3D printing	1960-1995
Higgs boson	1981 - 1988
Self-driving	1960-1998

Once we know the pre boom period we consider only articles from that period. There should be something in these articles that a human annotator would call weak signals. In this way we compile the signal corpus. To compile the non-signal corpus, which corresponds to the negative learning, we take the documents from pre boom period and any other articles from the pre boom period that do not exhibit any unusual distribution. We kept the same ratio between the number of signal vs. non signal documents as the one for the corpus described in Section 3.

By feeding into the deep learning algorithm the vectors representing these documents, our intuition, is that the more topics and the more documents we have from the pre boom period the higher accuracy we obtain. So we are going to test the accuracy of (i) the unsupervised learning and (ii) the dependency of accuracy on those parameters: number of topics, the detection accuracy of the pre boom period and the number of documents considered as weak signals. We test now on randomly chosen 1,500 documents out of which 500 are signals, and none of them contains any references to the topics we used to collect the training corpus. We obtain the following results (Table 8):

Table 8. Unsupervised Weak Signal Prediction.

	Weak Signal	No Signal
CR	0.45	0.89
LSTM	0.44	0.93

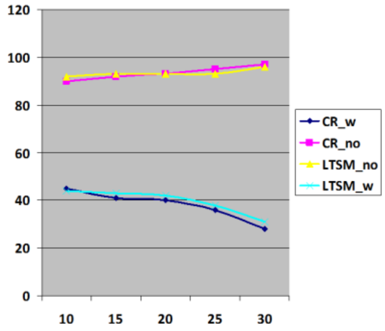


Fig. 9. Dependency on pre boom accuracy

The results considering 10 topics are very good. They are well above the *tf-idf* baseline and they are close to the performances of the SVM. We artificially varied the pre boom period in order to include more from the pre boom period, by increasing the boundaries by 10% to 30%. We notice a major deterioration in results; the accuracy was dropped to 0.3. By including more documents, the results did not go up in a statistical significant way. These findings suggest that pre boom epoch accurate detection plays a very important role, and that probably few thousands of documents, cumulatively; represent a sufficient statistics for this task. Little is gained by adding any other information, either in a supervised way or not.

Finally, we considered the following experiment to complete the picture. We tried to predict whether there will be a boom only in the distribution of the individual topics in the pre boom. That is, by just looking at the pre boom period we try to predict whether that respective topic will materialize in a fully-fledged trend. We considered the mean difference between the pre boom period and the pre boom period with an empirical threshold, the intuition being that a topic exhibiting a large difference; it is likely to witness a boom. The accuracy was less than .08, showing that the analysis of individual topics will tell us something about trends only when it is too late.

6 Conclusions

This study is an experiment on weak signals prediction. The possibility of trend prediction based on weak signals is very exciting and it has many applications. Our study shows that even when we do not know what the weak signals are, we are still able to use them in predicting future trends via deep learning methods with unsupervised learning. A carefully analysis of the experiments we carried out systematically from *tf-idf* to LSTM provide us with insights on choosing one type of learning over other possible candidates.

Next, we would like to experiment with other deep learning algorithms. A starting point is to understand better how we could narrow down the search for weak signals. The results suggest that we can have a major improvement of several points if we could pin point a paragraph instead of a document as source of weak signals. Therefore, our next effort is to narrow down the search for pre boom period at the paragraph level, rather than document level.

Acknowledgments: This project is funded by the Ministry of Research and Innovation within Program 1 – Development of the national RD system, Subprogram 1.2 – Institutional Performance – RDI excellence funding projects, Contract no.34PFE/19.10.2018”.

References

1. Amarandei, S., Fleșcan, A., Ioniță, G., Turcu, R., Trandabăț, D., Gîfu, D. Key Biomedical Concepts Extraction. In: D. Gîfu and D. Trandabăț (Eds.). Proceedings of the International Workshop on Curative Power of Medical Data, D. Gîfu and D. Trandabăț (Eds.). "Alexandru Ioan Cuza" University Publishing House, Iași, 21-26 (2017).

2. Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Martenson, C. and Svenson, P. Harvesting and Analysis of Weak Signals for Detecting Lone-Wolf Terrorists. Security Informatics 2, no. 11 (2013).
3. Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stock-Meyer. Alternation. Journal of the Association for Computing Machinery, 28(1):114-133 (1981).
4. Cohen, K., Johansson, F., Kaati, L., and Mork, J. Detecting Linguistic Markers for Radical Violence in Social Media. In: Terrorism and Political Violence 26, no. 1: 246-256 (2014).
5. Gerrish, S.M. and Blei, D.M. A language-based approach to measuring scholarly impact. In: Proceedings of International Conference of Machine Learning (2010).
6. Gifu, D. Contrastive diachronic study on Romanian language. In: Proceedings FOI-2015, S. Cojocaru, C. Gaindric (eds.), Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, 296-310 (2015).
7. Gusfield, D. Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK (1997).
8. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd ed., Springer (2008).
9. Mamishev, A.V. and Sargent, M. Creating Research and Scientific Documents Using Microsoft Word. Microsoft Press, Redmond, WA (2013).
10. Mamishev, A.V. and Williams, S.D. Technical Writing for Teams: The STREAM Tools Handbook. Wiley-IEEE Press, Hoboken, NJ (2010).
11. Mihalcea, R. and Nastase, V. Word Epoch Disambiguation: Finding How Words Change Over Time. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, vol. 2: 259-263 (2012).
12. Popescu, O. and Strapparava, C. Behind the Times: Detecting Epoch Changes using Large Corpora. International Joint Conference on Natural Language Processing, Nagoya, Japan, 347-355 (2013).
13. Popescu, O. and Strapparava, C. Time corpora: Epochs, opinions and changes. Knowledge-Based Systems, 69:3-13 (2014).
14. Abu-Mostafa, Y., Magdon-Ismail, M., Lin, H.-T.. Learning from Data, amlbook.com (2013).
15. Wang, X., Gerber, M.S., and Brown, D.E. Automatic Crime Prediction using Events Extracted from Twitter Posts. SBP, LNCS 7227:231-238 (2012).
16. Wang, X., McCallum, A. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. KDD 2006, USA (2006).
17. Wang, C., Blei, D.M., and Heckerman, D. Continuous Time Dynamic Topic Models. Proceedings of Uncertainty in Artificial Intelligence. Zunino, L., Bariviera, A. F., Guercio, M. B., Martinez, L. B., and Rosso, O. A.: On the efficiency of sovereign bond markets. Phys. A Stat. Mech. Appl., 391: 4342–4349 (2012).
18. Xin, Y.: Linear Regression Analysis: Theory and Computing (2009).