Tacit Knowledge - Weak Signal Detection

Alina Irimia, Paul Punguta, Radu Gheorghiu
firstname.lastname@uefiscdi.ro

Abstract

Before a certain topic becomes a very searched subject on news platform, there are some weak signals that, if correctly recognized and handled, may anticipated the popularity of that topic. One big problem with detecting such weak signals is that their recognition relies to a large extent on human tacit knowledge. Human tacit knowledge is a type of information having as main characteristics the fact that there is not a direct formal definition of it, and there is not a direct label in the text which explicitly marks it. In this paper we report on building an annotated news corpus for detection of weak signals. We also report on experiments using a supervised machine learning technique.

1. Introduction

In a diachronically ordered news corpus we can discover that a particular breakthrough event may have been predicted by corroborating small pieces of evidence existing in previously published pieces of news. That is, there are no pieces of text that directly mention or describe the breakthrough event, but there are scattered paragraphs, each one containing a faint and indirect indication to a certain possibility that further on becomes a breakthrough-event. Not being definable, this type of information cannot be identified by a precise set of rules written in a guideline for annotators. It is part of human tacit knowledge to identify the causes and consequences of certain events. In this paper we focus on weak signals, that is, on the information that a human reader is able to extract from a piece of news which is no more than a hint that a certain event is going to happen. The task we address is the classification of pieces of news into two categories: containing weak signals vs. non-containing weak signals. We have compiled a large corpus of news, made of some 40,000 scientific articles published in the last 50 years. A team of annotators were asked to annotate each piece of news as a whole according to whether the news contained or not weak signals. The annotation was carried out individually and conflicting opinions were discussed without any pressure to eventually reach a total agreement, via a process that is presented in details in Section 3. We selected a subset of roughly 20,000 documents on which the inter agreement was almost perfect (more than 99%) regarding the existence or non-existence of weak signals. We devised a set of machine learning experiments using this corpus. In section 4 we present the learning methods The fundamental result we report after these experiment is that machine learning methods can be used efficiently for tasks where the human tacit knowledge plays an important role. Few research directions which will investigate other aspects related to prediction and tacit knowledge are presented in the Conclusion and Further Research section.

2. Related Work

The literature on weak signals is not very large, as this field is about to emerge. A ground breaking paper (Brynielsson et al., 2013) was looking mainly at weak signals for detecting deviational behavior in order to efficiently provide preemptive counter measures. However, the probabilistic model presented is very close to the one used in language modeling, being 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 to predict classes of similar words for topics that are 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 the intention of the speaker. However, in scientific prediction this relationship is much more blurred, if it exists at all. We believe a new technology must be used. The diachronicity, that is the evolution of certain topics in mass media over time, is linked to detection of weak signals. Diachronicity is also an emerging field. We found useful two statistical tests presented for epoch detection in (Popescu and Strapparava, 2014), or temporal dynamics in (Wang and McCallum, 2006; Gerrish and Blei, 2010). In (Abu-Mostafa et al., 2012) we found very useful insights from dealing with discriminative analysis and support vector machine respectively. In order to improve our results we had to be able to deal with the masking effect and to understand how we could restrict further the objective function. The work of (Popescu and Strapparava, 2013; Popescu and Strapparava, 2014) is focused on diachronic analysis of text, in particular on trends. Their work centers on finding non-random changes in distribution of topics. However, their work is not concerned with prediction on the further evolution of the analyzed topics. In (Rocktäschel et al., 2015) the principle of an attentive neural network is presented. We used these principles to implement the network presented in Section 4.4. The literature on neural networks has become rich recently and there are more than a few papers reporting on their performances on semantic tasks, such as textual entailment, semantic text similarity, short text clustering (Mueller and Thyagarajan, 2016; Palangi et al., 2016; Xu et al., 2015). However, these approaches rely on the existence of a word or sentence level annotation, and an approach based on sequence to sequence alignment is doable. In this sense, our study extends these findings, by showing that it is possible to achieve good performance for tasks where there are no direct sequences of words that are aligned.

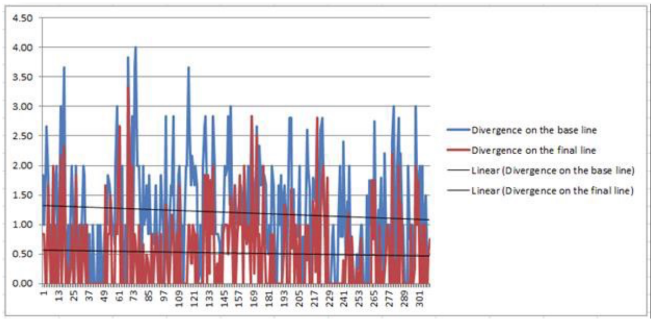

Figure 1: Towards reaching a stable shared tacit knowledge.

3. A weak Signal Corpus

Our assumption is that weak signals represent a form of tacit knowledge. As such, it may be counterproductive to define a formal set of guidelines aiming to precisely 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 the whole document. Most of the annotated paragraphs contained 100 to 250 words. Therefore, we obtained two annotated corpora, which, for convenience, we refer to as short and long respectively. The long corpora, LC, refers to full documents as training/test corpora. The short corpora, SC, refers to paragraphs. There is no perfect overlap between these two corpora; approximatively 15% of paragraphs come from different documents than the ones considered on LC corpora. The annotation is binary, *yes* or *no*, signaling the existence or lack of weak signals, respectively. In case of SC all the paragraphs that were not explicitly classified as *yes*from the analyzed documents are 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 mis-classification. Eventually we have the following distribution in SC, LC corpora, see Table 1:

	Weak Signal	No Weak Signal
LC	4,100	14,020
SC	3,700	14,500

Table 1: Weak Signal Corpus

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. For these documents there was a large agreement regarding their category, over 99% agreement. The exact process of annotation is described below.

3.1. Annotation via tacit knowledge

We had a team of 18 undergrad volunteers. The main question was whether the periphery of tacit knowledge will become stable after practicing several hundred annotations or whether there would be a large area subject to dis-agreement between annotators. On a given set of 300 documents the annotators were encouraged to discuss their doubts and to defend their position in case of disagreement. 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 also 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. However, the average time for making a decision decreased for time between these two experiments, see Fig.2. It can be considered that these results suggest that this task, in spite of being 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 a grey zone where experience does not help. This behaviour tends to help an automatic classifier, as it does not have to be very precise in order to obtain a human like performance. After a preliminary round of trial annotation of several hundreds of documents, we decided to create a taxonomy that sprung naturally from this experiment. This flat taxonomy has the following components:, technology, innovation in services, trend shift, behavioral change, major actor move, breakthrough discovery, top research, wild card.

The intention in using these labels was to try to capture the intuition of annotator on why a certain document/paragraph is considered as carrier of weak signals. As people usually tend to overweigh the famous research centers, famous names etc, this taxonomy helps us to see if there are indeed any subjective differences that may affect the learning process. The indication here was that wild card, which is

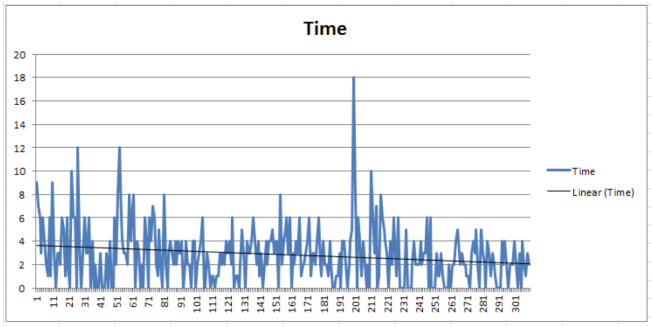

Figure 2: Average time for making a decision.

	Technology	Others									
Categories		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	

Figure 3: Weak Signal taxonomy Distribution

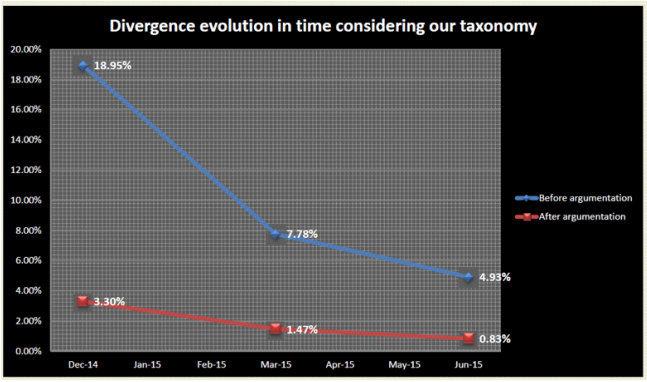

Fig 4. Reaching consensus over taxonomies

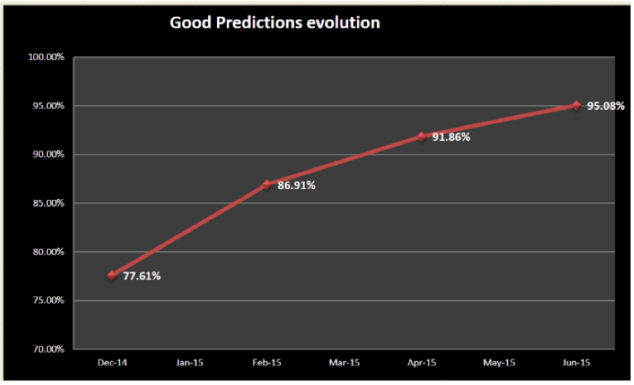

Fig 5. Control Group Judgement

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

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

equivalent to none of the above, is always a valid option.

It came as a surprise that the annotators did not want to use often the wild card taxonomy. It can be seen that 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. We wanted to check whether this consensus was reached due to an increasingly strong and commonly shared tacit knowledge, that is, due to acquiring an expertise, or due to accepting a dominant view.

A control group checked the validity of the agreement and what we found is that the results strongly suggest the first alternative, that is acquiring an expertise, see Fig. 5.

In conclusion, all these experiments strongly suggest that we have a tacit knowledge about weak signals that is shared at least 80

4. Learning Weak Signals

In this section we present a series of learning approaches which we tried step by step.

In a supervised approach, finding the pieces of news containing weak signals is a binary classification task. A first approach is to use tfidf 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 on the basis of which a 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 from scikit library, and support vector machine, linear SVM form Weka library, respectively. See also the equations 1 and 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. In fact, we will see in the next section that the gradient descent is an effective technique for this task. At this point, we have to understand whether the projection of the data into a bi-dimensional space will lead to con-like structures, that is, that the data can be 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. 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 SC and 0.12 for LC respectively. As we can see both QDA and SVM scored significantly better than that. And indeed there is a non-random difference between QDA and SVM results.

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 Table 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, eq. 1, it 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 dimension is unknown, on the basis of the Tables 2, 3 it is intuitively tempting to consider that the VC bound is indeed too loose for this task. As such, we can do better in estimating the posterior probability. The right question is whether we have enough data to train a more detailed clas-

	Weak Signal	No Signal		Weak Signal	No Signal
QDAcr	0.412	0.877	QDAcr	0.38	0.901
SVMcr	0.663	0.913	SVMcr	0.472	0.946
QDAts	0.403	0.865	QDAts	0.365	0.890
SVMts	0.610	0.905	SVMts	0.455	0.930

Table 2: Supervised Learnig of Weak Signals

sifier. We may guess that deep learning methods may be up to the task.

5. Conclusion

In this paper we presented an experiment on prediction. Rather than final, we consider these results as a very promising beginning for research into this field. The possibility of trend prediction on the basis of weak signals is very exciting and it has a lot of 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 supervised learning. This is an excellent result, showing that it is viable to talk about predictions. In many applications, it becomes critical to have an accurate prediction. In science, making predictions almost equates to having a bright idea on how apparently disparate small achievements may converge to a breakthrough discovery. In our digital era, accessing billions of documents is easy but selecting the ones carrying relevant information which is not yet fully developed is difficult. 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. So our next effort is to narrow down the search for the pre boom period at the paragraph level, rather than document level.

6. References

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook New York, NY, USA:.

Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Mårtenson, C., and Svenson, P. (2013). Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, 2(1):11.

Gerrish, S. and Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *ICML*, volume 10, pages 375–382. Citeseer.

Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., and Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):694–707.

Popescu, O. and Strapparava, C. (2013). Behind the times: Detecting epoch changes using large corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 347–355.

Popescu, O. and Strapparava, C. (2014). Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3–13.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Wang, X. and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM.

Wang, X., Gerber, M. S., and Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer.

Xu, J., Wang, P., Tian, G., Xu, B., Zhao, J., Wang, F., and Hao, H. (2015). Short text clustering via convolutional neural networks. In *VS@ HLT-NAACL*, pages 62–69.