Automatic Prediction of Future Business Conditions

Lucia Noce

L. Noce, A. Zamberletti, I. Gallo, G. Piccoli, and J. A. Rodriguez
Department of Theoretical and Applied Science (DiSTA)
University of Insubria

PolTAL

17-19 Septemnber 2014

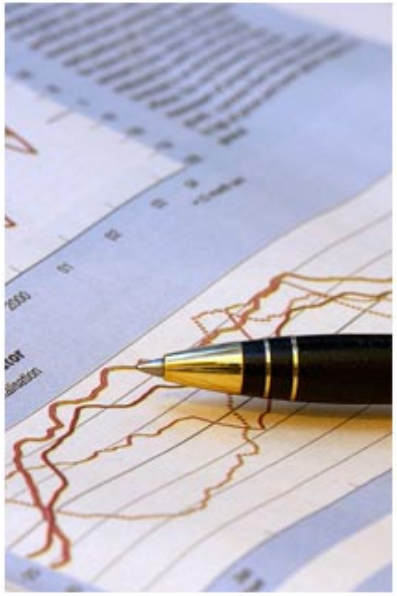
Goals:

- Help business analysts and policy makers to more accurately forecast firm's future behavior and performance.
- Automatically extract forward looking-statement from a specific type of formal corporate documents called earning call transcripts.
- Exploiting both Natural Language Processing and Machine Learning techniques.

Introduction - What is an earning call transcript?

An earning call transcript is the verbatim textual record of a conference call between the management of a public company, analysts, investors and/or the media to discuss the financial results the firm achieved during a specific past reporting period (e.g., quarter, fiscal year).

Firms provide a substantial amount of future-related information.



Introduction - Why earning calls (E.C.)

Why earning calls:

- The narrative/qualitative information enclosed in earning call transcripts can be successfully exploited to predict both short term firm-specific performance and future macroeconomics fluctuations.

- E.C. are formal and well-formed documents.

Introduction - Forward-looking statement

E.C. are used to extract forward-looking statements:

- A forward-looking statement is defined as a short sentence that contains information refer to a direct effect in the future (e.g., plans, predictions, forecasting, expectations and intentions).
- **OUR GOAL:** automatically extract future-looking statements from generic earning call documents.

Classical text classification problem

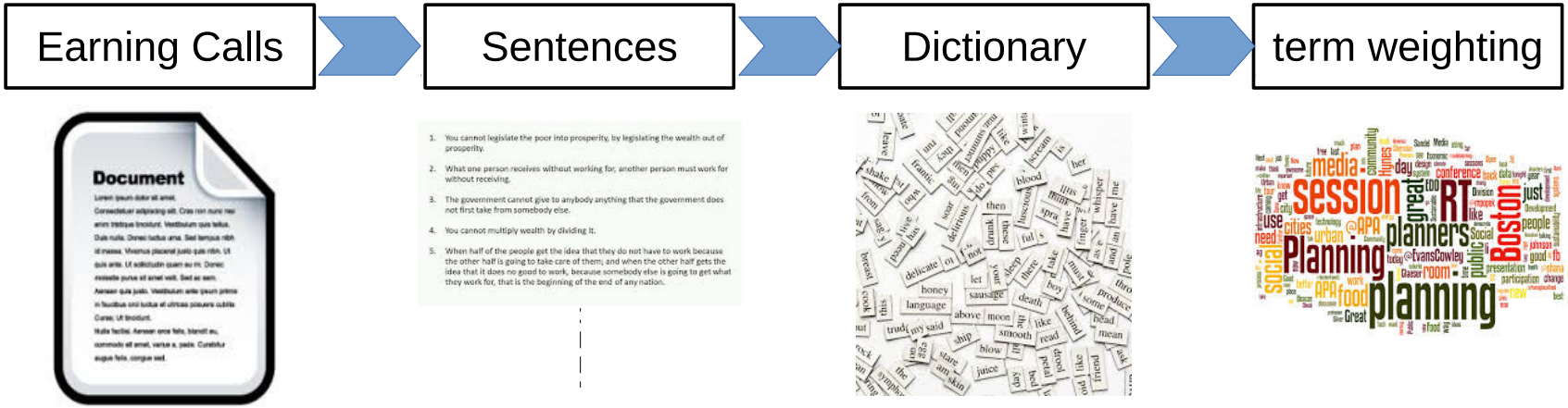
The problem has been approached as follows:

1. Given a finite non-empty set of documents $X = x_0, ..., x_n$ and a set of categories $\Omega = \omega_1, ..., \omega_c$, the proposed task requires to assign to every pair $(x_i, y_j) \in X \times \Omega$ a boolean label yes/no
2. Each document $x_i \in X$ is usually represented as a vector $\phi(x_i)$ in which each element measures the number of times that a given term in x_i, contained into a finite dictionary of terms D, appears in x_i.
3. Given $x_i \in X$ and a dictionary of terms $D = t_1, ..., t_d$, the document is represented as $\phi(x_i) = (x_i, ..., x_{d_i})$,where each $x_{j_i} \in \phi(x_i)$ measures the frequency of occurrence of the term $t_j \in D$ in the document $x_i \in X$.

In our method:

- The set of categories $\Omega = \{\text{FLS}, \text{NFLS}\}$.
- The set of documents X is composed by all the sentences of all the earning calls used.
- The dictionary of terms D is build using the $Relevance$ formula.
- The feature vectors $\phi(x_i)$ are provided as input to a supervised classifier.

Dictionary



From left to right:

- The starting earning call documents is split into sentences.
- The sentences are split into words to build a dictionary.
- To each term in the dictionary a weight is assigned in order to detect a small subset of terms that are characteristics of forward-looking sentences.

The weight for each term is computed using this weighting formula:

Penas, A., Verdejo, F., Gonzalo, J.: Corpus-based terminology extraction applied to information access. In: Proceedings of Corpus Linguistics. CL (2001)

$$Relevance(t_i, sc, gc) = 1 - \frac{1}{\log_2 \left(2 + \frac{F_{t_i,sc} \cdot D_{t_i,sc}}{F_{t_i,gt}}\right)}$$

Where:

1. sc is the specific corpus, it corresponds to the subset of sentences from X, that were tagged by experts as FLS.
2. gc is the generic corpus, it is composed by the whole set of sentences X.
3. $F_{t_i,sc}$ is the relative frequency of the term $t_i \in D'$ in the specific corpus sc.
4. $F_{t_i,gc}$ is the relative frequency of the same term $t_i \in D'$ in the generic corpus gc.
5. $D_{t,sc}$ is the relative number of documents of sc in which the term $t_i \in D'$ appears.

Final dictionary

Once the relevance of every term $t_i \in D'$ has been computed, it is possible to obtain the final smaller dictionary D:

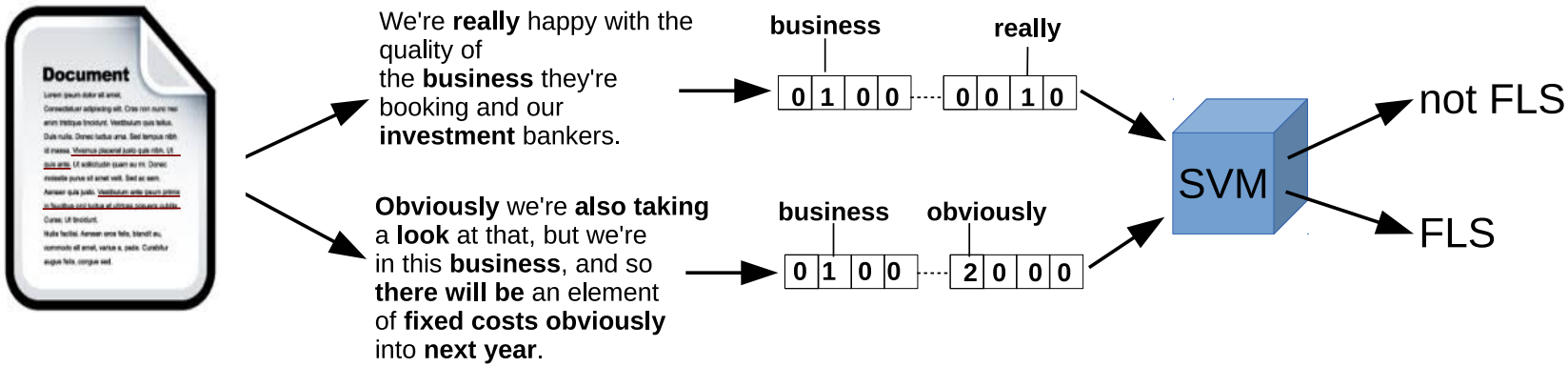
- Only the terms having the highest weights appear in the final dictionary.
- The final smaller dictionary D is obtined by removing all those terms whose relevance value is lower than a threshold ψ, as follows:

$$D = \{t_i \in \mathcal{D}' \mid Relevance(t_i, sc, gc) > \psi\}$$

We decided to employ a Support Vector Machine (SVM) classifier:

- The same model was used in most of the previous related works.
- It can lead to optimal results with minimal tuning effort.

An example: how the feature vectors are generated



From left to right:

- A vectorial representation using the final weighted dictionary D is given to each sentence.
- Those feature vectors are finally provided as input to the SVM classifier.
- The SVM classifier assigns them a label denoting wether they represent forward-looking statements or not.

We downloaded the earning call documents provided by three leading firms:

- International Business Machines Corporation (IBM)
- Exxon Mobil Corp (XOM)
- J.P. Morgan Chase & Co. (JPM)

For each firm, we picked the E.C. for third quarter of the years 2013, 2012 and 2011.

Dataset: how do we make it.

The ground truth:

1. The earning call documents were split into sentences using a classic sentence detector.
2. Those sentences were then manually labeled by a team of experts in economic and finance.
3. Each sentence has been processed multiple times by different experts.
4. The degree of inter-rater reliability among the experts is equal to 89.73%.

Experimets: train and test sets

Train and test:

- Total amount of 3148 tagged sentences.
- We split the dataset into train and test sets following a classic $\frac{2}{3}$-train, $\frac{1}{3}$-test split rule.
- The sentences from the three firms were split among the train and test sets in order to make the dataset as heterogeneous as possible and independent from any firms' specific language style.
- SVM classifier model was trained using 5-fold cross validation.

Results: proposed method

Accuracies achieved by the proposed method while varying the value of the threshold parameter ψ.

threshold ψ	OA (%)
0.2	87.57
0.3	87.57
0.4	83.69
0.5	83.16

High values of ψ overprune the dictionary by removing terms that are highly discriminative for forward-looking statements.

Different approaches: BOW

Classical Bag Of Words approach:

- Dictionary composed by all the words from the sentences in the training set.
- The sentences from the train and test sets were then represented as BOW feature vectors.

Different approaches: PoS

Part of Speech tagger:

- PoS tagger algorithm is used to build a dictionary of tags that is used to build the feature vectors.
- The number of possible tags assigned to each word by the PoS tagger is equal to 42.

Experiments: parameters

Parameters:

- Varying the size of the dictionary of terms.
- Varying the ***ngram*** taken into account during the building phase of the dictionary.

  In detail, when using an ngram value of 1, we build D using the single words extracted from the sentences in X. On the other hand, when using an ngram value of 2, we not only consider the single words but also all the possible sequences of two consecutives words inside each sentence.
- Varying the **context** considered during the building phase of the feature vectors.

  Given a sentence $x_i \in X$ in position r within the earning call EC to which it belongs, when context is taken into account, we build the feature vector for x_i by summing the contributes from the sentences in positions $r-1$, r and $r+1$ within EC.

We reproduced and tested the approach proposed by Bozanic *et al.*

Bozanic, Z., Roulstone, D.T., Van Buskirk, A.: Management earnings forecasts and forward-looking statements. (2013)

- Recent (2013) work from the accounting and finance literature.
- A sentence is labeled as FLS iff it containes at least one forward-looking terms.
- The list of all the forward-looking terms used is provided by authors.

Results

Comparison between the results achieved by the proposed method and other approaches analized for the dataset measured using the metrics: *Overall Accuracy* (OA), *Precision* (p), *Recall* (r) and *F-measure* (f).

method	OA (%)	p (%)	r (%)	f (%)
Bozanic	35.56	20.05	78.21	31.91
BOW	84.93	69.37	48.94	57.39
PoS	81.03	69.10	48.07	56.69
Proposed	**87.57**	**74.82**	**53.23**	**62.21**

Thank **You!**

Department of Theoretical and Applied Science (DiSTA)
University of Insubria
Varese, Italy

email: lucia.noce@uninsubria.it

web: http://artelab.dista.uninsubria.it/