Sentiment Analysis on Economic Texts

Claudiu-Cristian Musat, Stefan Trausan-Matu

"Politehnica" University of Bucharest,

Splaiul Independentei 313,

Bucharest, Romania

claudiu.musat@gmail.com, trausan@cs.pub.ro

Abstract. We propose a new approach to opinion mining that relies not on explicit occurrences of opinions in the given texts but on the polarity of the chosen topic. We apply this approach on economic texts extracted either from weblogs or financial publications. We aim to prove that the vast majority of economic texts contain personal opinions even if explicit opinion indicators are missing. We then intend to prove that documents with a given polarity are more likely to be interlinked with others having the same polarity than with documents having the opposite one.

Keywords: Document Polarity, Economy, Opinion Mining, Contextual Network Graph

1 Introduction

One of the definitions[1] of Opinion Mining states that given a set of documents that contain opinions about an object, opinion mining aims to extract attributes of the object that have been commented on in each document and to determine whether the comments are positive, negative or neutral.
But there are situations where we need to broaden this perspective in order to detect opinions that only derive from context and are not expressed directly. We limit the search for such opinions to economic texts.

1.1 Contextual Network Graphs

In order to limit the analysis on economic texts alone, we need to filter out texts that have no link to the economic environment, which translates to classifying input texts as economic or not. We used a modified version of a Contextual Network Graph (CNG) [2] for that task. Introduced by Ceglowski et. al. in 2003, CNG's are a powerful tool for mining unstructured data and a good competitor for LSI [3]. In short, the method uses a term-document matrix to create a bipartite graph that

represents the entire collection. The graph can then be queried by distributing energy from one or more of its nodes and the results are those nodes that receive energy above a predefined threshold.
In our case, we need the graph to indicate whether a certain term is a good indicator of an economic text.
By having only known economic texts in the graph we can check whether by energizing it from a node representing the search term it will activate known economic terms. Doing so repeatedly increases the chances of a good recall.

1.2 Opinion Mining

Opinion Mining, or sentiment analysis is a fast developing research area, including topics and methods from text mining, natural language processing and computational linguistics. Widely investigated since 2005, this research area can be discussed from multiple viewpoints such as information sources, methodologies or the objectives themselves.
Generally the aim of the processes involved in sentiment analysis is to determine the orientation or polarity of expressed opinions. Different levels of subjective information can be found at the word, sentence or document levels.

One of the first works in the field was that of Pang et. al. [4], where documents were classified according to the overall opinions expressed. Then Wiebe et. al. [5] discriminated between subjective and objective sentences. Later the focus shifted to determining the polarity of expressed opinions, ranging from a positive/negative dichotomy to more complex classifications based on opinion strength [6].
The next target of opinion mining was to link the opinions to concepts extracted from texts. More recently the problem was broken down even more with the introduction of the faceted opinion concept[7].The goal of this attempt was to determine precisely what aspects of the concepts the expressed opinions should be linked to.
However all the methods above rely on the presence of explicit opinions in the analyzed texts, from the obvious (e.g. a horrifying experiment) to the more refined. That is perhaps one of the main reasons why opinion mining has been limited in many cases to extracting opinions about products such as films[4], where client feedback is direct (e.g. liked/disliked the film).
There are many cases where opinions are expressed indirectly and where the use of methods that work well on film reviews might prove dissatisfying. The lack of explicit opinions does not always imply the lack of an overall opinion, and that is an area that has received comparatively little attention so far.

The paper continues by stating the problem we are addressing in Section 2. The proposed system follows in Section 3 and we conclude in Section 4.

2 Opinions in Social Sciences

Although the vast majority of specialty texts in social sciences claim to be unbiased, with the author merely being an outside observer, that is rarely the case. Even in a closed environment there can be multiple viewpoints and the results obtained by using them can vary. But in the real world, as both the number of observations and variables increase dramatically, number of possible perspectives increases as well.

A physician can rigorously express his opinion about a case and argue it with hard evidence saying for instance that the problems the patient is facing are severe. However a different doctor can examine the same patient regarding a different bodily function and correctly state that everything is going well. The two doctors' opinions regarding the same patient can be opposite and still they can both be right, the only thing that differs being their perspectives.

Economy is a social science and the same applies here. In the economic environment today the main focus is usually on aspects of the unfolding crisis. If in the fall of 2008, after the fall of Lehman Brothers almost everybody on the planet agreed that we were all in the midst of a very serious situation and that the outlook was bad. Nowadays traders in London and on Wall Street generally argue the crisis is over or at least closing to an end while academic economists such as Nobel Prize Laureate Paul Krugman or Professor Nouriel Roubini argue that this might be just a bear market rally (an increase in stock market valuations that follows a sharp earlier decline and that will be followed by another sharper one in the future).
The reason for such dissonance is that each economist watches a subset of all existing economic indicators he or she considers relevant. The traders in the New York and London financial centers are more inclined to watch the evolution of the stock market while academics are more likely to place their trust in higher level macroeconomic indicators such as the Baltic Dry Index [8], that monitors the total amount of goods shipped from one place to another, hence a good indicator of the state of world industrial commerce.

Hedge fund managers are another type of market participant that do not fall in either of the previously mentioned categories. Given that they are entrusted with the fortunes of many they must draw conclusions from all available data. Yet, as Ambrose Pritchard Evans so obviously states it [9], their conclusions differ heavily. As Bohr put it nearly a century ago, prediction is difficult, especially of the future. Critics may argue that the case where the same situation can be at the same time positive for one person and negative for another is quite common. But when the problem comes down to buying or selling the same securities, that is hardly the case.

A world apart are the economic journalists and independent bloggers who can either express opinions directly, forecasting a certain outcome for economic processes, such as the beginning of an economic recovery, or they can choose to present factual data. The text in [9] is relevant in another way as well – the author himself takes a stance in spite of presenting two antagonistic opinions from two renowned fund managers, by selecting one of the opinions as the post title (coincidentally or not, that opinion is the same expressed in previous articles by the same journalist). Thus an otherwise perfectly objective text has a strong bias.

There is one thing that connects all economists or economically inclined writers. They all have personal opinions that emerge in their writings even if it is their desire to be objective. Their opinions are unmasked by the positive or negative aspects of the economic life that they choose to describe.

3 Proposed System

The system we are currently developing aims at extracting opinions from economic texts, even if at first glance they seem objective. It comprises three main parts – a web crawler that extracts data starting from the given websites, a topic detection module based on a modified CNG and an opinion extraction module.

In short, starting from a list of known financial weblogs and newspaper articles, we first obtain the contents of the webpage, check whether it is a relevant document and, if so, we add the documents linked from that webpage to the list of pages to be processed. For each relevant document we then determine its polarity.

3.1 The Web Spider

The crawler has the task to get the whole text of a web page and determine the relevant text within. Although the task seems straightforward, there are problems that need to be overcome.

We start by defining the relevant text we want to extract depending on the type of webpage we are extracting it from. Thus, for a blog post it is the entire text of the post, leaving aside the comments and other data that we might find. The same holds for a newspaper article. But for obtaining the text from the home page of a blog we need to concatenate the contents of all the abstracts of the posts present there.

We used the LWP::Simple Perl module[10] to obtain the contents of the web pages and the HTML::Parser [11] module to parse the returned text. We then created another Perl module to only return the text with the highest likelihood to be the actual contents of the blog post. We also used a threshold for the document's length to filter out small and likely less relevant posts.

3.2 Text Processing

From our experiments, every processed document links, on average, to more than sixty other documents that pass the conditions above. In the face of such an exponential increase in the number of documents that need processing, we singled out those that fall in the economic category and refer to known economic indicators such as confidence, market or activity indexes or more generally to the current economic crisis.

Also, because of this multitude of documents to be checked, we could not use a standard CNG because its size grew too large and the time to process a new document

became prohibitive. We used a two step approach instead. First the extraction of new keywords from the analyzed documents and then the categorization itself.

3.2.1 Keyword Identification

We need to increase the initial set of expressions with others possibly relevant ones from the documents we analyze, in other words – to identify new keywords.

We initialize the graph with a set of words and expressions relevant to the economic community, the initial keywords, K_I. Then, for every document we obtain the list of words and known bigrams and trigrams that remain after stemming and stop word elimination. Stemming and stop word elimination have been done using Perl's Lingua::Stem[12] and Lingua::Stopwords[13] modules. Let us denote the document's selected words and expressions W_D.

We add the current document D to the graph and then add all its extracted words W_D if they haven't been added along with a previous document.

The latter part is to energize the graph from all of the items in W_D and check which keywords in K_I receive sufficient energy to exceed a predefined threshold. For each of the expressions in W_D we add the energy that from them reached the keywords in K_I.

The outcome of this process is based on the fact that expressions relevant to a situation co-occur. The results have been encouraging. For example, we found that "subprime" and "prime" are highly linked to "loan" and that "chilling" is linked to "CHF" – the Swiss Franc and to "foreclosure" – the repossessing of homes by banks.

The method needs some improvements though, because a large number of words were linked to important names in economics – such as Bloomberg, which probably shouldn't have been in the initial subset in the first place.

After sufficient documents have been processed and the number of keywords is satisfying we can start classifying the documents themselves.

3.2.2 Text Categorization

We reduced this second phase of the text processing to a simple keyword identification problem. If enough of the previously detected keywords are present within a document, we can assume with a high degree of confidence that it has important links to the economic sector.

3.3 Polarity Identification

Polarity Identification is the key element of the system. As we stated in the introductory part, the goal is not to find expressions of opinions in the analyzed text, but rather to find indirect opinions, that derive from the topic discussed.

To do that we need two pools of terms. The first is that of economic indicators and the second is a collection of terms that indicate growth or decrease. The first pool must then be divided into two subsets. One that contains indices that grow along with the economy as a whole and one with indices that shrink during a period of economic boom. For instance “stocks” and “industrial output” grow when the economy is doing well but prices for “corporate bonds” decrease in the same period. We will call the first “positive” indicators, and the latter “negative”.

For the second pool of terms we need more than just synonyms and antonyms for growth. Our approach is to search the WordNet glosses for words that indicate growth or rather the opposite. Each word that has in its gloss a member of the initially chosen synsets will be signaled for validation.

Having constructed these two pools of terms we will have the necessary data to look for co-occurrences of terms from the first set – that of economic indicators and the second – that of growth terms. We are only interested in the case where members of the two sets appear in pairs – for example “the economy is growing” where “economy” (its seed actually) will be a member of the first set and “grow” a member of the second.

A document in which the majority of pairs indicate the growth of a positive indicator or the decrease of a negative one will be labeled as having a positive polarity, whereas a document in which the majority of pairs indicate the growth of a negative indicator or the decrease of a positive one will have a negative polarity.

3.4 Link Extraction and Connection Graph Creation

After parsing the web page text, for the selected ones we needed to extract the outgoing links. Here too we used a Perl module: HTML::Linkextor[14] to obtain the list of links within a given page. After having the list we need to filter out links to scripts, images and most importantly to previously visited web pages. Not doing so results in an infinite loop.

The resulting list is then added to the original list of links to be processed, creating a breadth first search. We realize that unless we are extremely lucky and choose an initial set of links that have a small number of outgoing links of their own, the process will never end- especially if we keep in mind that the Internet itself is growing ever faster. That is why we need to output preliminary results whenever a threshold of recently processed documents is reached.

The connection graph will contain all the links between previously processed documents. Every time we process a link we add the edges between it and its own outgoing links and also count the number of links between documents with the same or different polarity. We hope that when the process is finalized we will be able to answer the question whether it is significantly more likely for two economic documents with the same polarity to be interlinked than if they had opposite polarity.

3.5 Preliminary results

Although more than two million web pages have been queued for analysis and roughly 280.000 have been downloaded and split into economic and not economic following the criteria above, a significantly smaller number have been categorized into objective and subjective. This is due to the need to, at least in the first stages of the ongoing research, manually validate the results. Thus, for the first hundred economic texts that have been labeled by a human expert, the system correctly marked 67 as either subjectively positive or negative. No real positive texts have been marked as negative and no real negative texts as positive.

4 Conclusions

By switching from the detection of explicit opinions in economic texts to the quantification of occurrences of predictions of future economic outcomes we hope to create a novel tool for opinion mining that is complementary to existing ones. Current results show that keyword identification and document classification work, and we hope to have similar outcomes for the latter phases – polarity identification and document polarity graph creation. If that should be the case, it will be easier to navigate through the sea of economic predictions by at least separating antagonistic ones.

Acknowledgements
This work is supported by the European Union Grant POSDRU/6/1.5/S/19 7713

References

1. Liu B.: Opinion Mining. WWW-2008, Beijing, 2008
2. Ceglowski M., Coburn A., Cuadrado J.: Semantic Search of Unstructured Data using Contextual Network Graphs. 2003
3. Cherubini M., Dillenbourg P., Viscanti L. ,:Mobile Search on Ubiquitous Collaborative Annotations of Space. FWS, Barcelona 2006
4. Pang, B., Lee L. : “Opinion Mining and Sentiment Analysis”, FTIR, 2008
5. Wiebe , J., Wilson T., Cardie C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210
6. Chen L.S., Chiu H.J.: Developing a Neural Network based Index For Sentiment Classification. IAENG Hong Kong 2009.
7. Mei Q., Ling X., Wondra M., Su H., Zhai C.: Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. WWW Calgary 2007.
8. Baltic Dry Index Evolution

http://www.investmenttools.com/futures/bdi_baltic_dry_index.htm
9. "RBS uber-bear issues fresh alert on global stock markets", Prichard-Evans The Telegraph, August 2009
10. Perl Module LWP::Simple http://search.cpan.org/~gaas/libwww-perl-5.831/lib/LWP/Simple.pm
11. Perl Module HTML::Parser http://search.cpan.org/dist/HTML-Parser/
12. Perl Module Lingua::Stem http://search.cpan.org/~snowhare/Lingua-Stem-0.83/lib/Lingua/Stem.pod
13. Perl Module Lingua::Stopwords http://search.cpan.org/~creamyg/Lingua-StopWords-0.09/lib/Lingua/StopWords.pm
14. Perl Module HTML::LinkExtor http://search.cpan.org/~zigorou/MozRepl-Plugin-LinkTools-0.01/lib/MozRepl/Plugin/LinkExtor.pm