[Proceeding] Supporting stock trading in multiple foreign markets: a multilingual news summarization approach

Original Citation:

Availability:
This version is available at: http://porto.polito.it/2656998/ since: November 2016

Publisher:
ACM

Published version:
DOI: 10.1145/2951894

Terms of use:
This article is made available under terms and conditions applicable to Open Access Policy Article ("Public - All rights reserved"), as described at http://porto.polito.it/terms_and_conditions.html

Porto, the institutional repository of the Politecnico di Torino, is provided by the University Library and the IT-Services. The aim is to enable open access to all the world. Please share with us how this access benefits you. Your story matters.

(Article begins on next page)
Modeling correlations among air pollution-related data through generalized association rules

Elena Baralis, Luca Cagliero, Tania Cerquitelli

Dipartimento di Automatica e Informatica,
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
E-mail: luca.cagliero@polito.it

Keywords Financial data mining; text mining; text summarization.

Abstract In today’s world, plenty of textual news on stock markets written in different languages are available for traders, financial promoters, and private investors. However, their potential in supporting trading in multiple foreign markets is limited by the large volume of the textual corpora, which is practically unmanageable for manual inspection. Although, text mining and information retrieval techniques allow the automatic generation of interesting summaries from document collections, the study and application of multilingual summarization algorithms to financial news is still an open research problem. This paper addresses the summarization of collections of financial documents written in different languages to enhance the financial actor’s awareness of foreign markets. Specifically, the proposed mining system (i) is able to cope with news written in multiple languages, (ii) generates multiple-level summaries covering specific and high-level concepts in separate sections, on behalf of users with different skill levels, and (iii) ranks the summary content based on both objective and subjective quality indices. These features are taking an increasingly important role in financial data summarization. As a case study, a preliminary implementation of the proposed system has been presented and validated on real multilingual news ranging over stocks of different markets. The preliminary results show the effectiveness and usability of the proposed approach.
1 Introduction

With the advent of the Internet, private and public financial investors have been overwhelmed by a continuous flood of textual news about finance and stock markets. News often cover the fundamentals or the analytics about stocks (e.g. outlooks, rating agencies’ feedback, monthly reports). Financial news are generated worldwide from heterogeneous sources and they are written in different languages. Exploring multilingual news on financial markets allow investors to consider a larger variety of assets spread all over the world and, thus, to identify more profitable assets to trade. Although daily and historical news on foreign markets are easily accessible through the Web, the knowledge of investors on foreign assets is often limited, because the manual exploration of this multilingual news corpora is extremely time consuming and, most of the times, practically unmanageable. This prompts the need for automating the process of knowledge discovery from financial news by means of advanced data analytics tools.

Sentence-based news article summarization Document Understanding Conference [2004] is the task of automatically extracting succinct descriptions of potentially large collections of textual news which consist of the most informative news sentences. The selected sentences are combined into a readable and promptly usable summary, which covers all the salient news facets. State-of-the-art summarization approaches often rely on data mining or information retrieval techniques (e.g. Baralis et al. [2012], Steinberger et al. [2011], Wang et al. [2011]). To perform statistics-based analyses, most algorithms assume that the input news all range over the same topic Document Understanding Conference [2004].

To the best of our knowledge, few attempts to analyze financial news by means of summarization algorithms have already been made Filipova et al. [2009], Otterbacher et al. [2006], Yang and Wang [2003]. To facilitate short-term trading activities, in Filipova et al. [2009] the authors generated summaries of news ranging over specific companies by using established similarity measures Tan et al. [2005]. Summaries on a given stock were generated in response to textual user-generated queries, which are automatically expanded with related terms by using the term frequency-inverse document frequency statistics Lin and Hovy [2003]. Hence, the context in which the underlying company operates is described by means of a uncontrolled vocabulary inferred from the news corpora and the resulting summaries combine specific stock information with more general contextual knowledge. In Otterbacher et al. [2006], Yang and Wang [2003] the authors applied fractal summarization strategies to reduce the computational load in summarization thus making the system portable to mobile devices. Similar to Filipova et al. [2009], the approaches proposed in Otterbacher et al. [2006], Yang and Wang [2003] have two main limitations: (i) they are not designed to cope with non-English-written news corpora, and (ii) they model the context of a stock as a set of keywords to be considered during summary generation.

In this paper we argue toward a new data mining-oriented system aimed at enhancing the awareness of different actors (e.g. traders, brokers, hedge funds) through the automatic generation of compact summaries of textual news on stock markets written in different languages. The proposed system, named Stock News Summarizer (SNS), automatically selects a potentially interesting yet easily manageable subset of news sentences. These sentences, which are ranked in order of decreasing relevance, can be manually explored in place of the potentially large multilingual news corpora. Sentence ranking can be driven either by objective (statistics-based) evaluators or by subjective metrics according to the expectations of the domain expert. Users can browse the summaries and provide feedback useful for driving future data analyses. The tool is multilingual because it is capable of summarizing collections of news written in different languages. To the best of our knowledge, the generation of multiple-level and multi-lingual summaries of financial news has never been addressed in literature.

A preliminary implementation of the SNS system was developed and experimentally evaluated on real multilingual news crawled from YahooFinance and GoogleNews and ranging over stocks of the American SP-500 and Italian FTSE MIB-40 indices. The achieved results demonstrated the usability and effectiveness of the proposed approach.

This paper is organized as follows. Section 2 introduces the addressed research issues, Section 3 presents the architecture of the SNS system, Section 4 summarizes the preliminary experimental results achieved during our study, while Section 5 draws conclusions and discusses future works.

2 Vision and challenges

Given a collection of textual financial documents (e.g. news, technical reports, fundamental analyses) written in different languages, the main research goal of financial text summarization is to study and design data mining approaches that yield compact yet informative summaries of the source documents, possibly tailored to different users or scopes.

This work addresses three related research issues that are currently open and, thus, potentially of interest for the research community:

1. What is the most appropriate abstraction level of the summary content? Multiple-level summaries are summaries that analyze a given topic at multiple abstraction levels. The system proposed in the work...
generates multiple-level summaries of financial news written in different languages. To allow experts to enhance their awareness of foreign markets, these summaries not only provide more insights into the underlying companies but also they give a more general overview of the corresponding financial sectors.

2. How to cope with financial news written in different languages? According to the notation used in document summarization Tan et al. [2005], multilingual systems are tools that are able to cope with collections of documents all written in the same language within the same collection but potentially different within different collections. They are a simplified version of cross-lingual systems, where a mix of documents written in different languages can be handled at the same time. The system presented in this study can handle multilingual collections of financial news, because it applies data preparation and summarization techniques that are portable to documents written in different languages. We consider the extension of the proposed system towards a cross-lingual environment as an open research issue, which we plan to address as future extension of this work.

3. How to select the most significant sentences? In document summarization sentence evaluation is the task of defining appropriate measures to judge the relative importance of the document sentences. In this paper we propose to use not only objective quality measures but also subjective ones. Subjective sentence evaluation takes advantage of the user feedback to properly decide which content is worth including in the summary. In the context of financial data analysis, we deem the integration of subjective metrics as particularly promising, because domain experts can give insightful interpretations of the financial news content.

3 Stock News Summarizer

The components of the envisioned architecture, as well as the interactions between such components, are shown in Figure 1. Hereafter we will refer to this architecture as the Stock News Summarizer (SNS). We envision a multilingual engine able to produce easily manageable summaries of news written in different languages, through which final users can enhance their awareness of foreign markets. SNS will support the decision-making process of financial actors by providing a clear and distinct picture of financial markets. To this aim, it generates multiple-level text summaries, which consist of (i) a summary of the information about specific financial assets (e.g. stocks), and (ii) a summary of the contextual information surrounding financial assets (e.g. the financial sectors, the market indices). Multiple-level content is useful for analyzing financial data from different viewpoints, on behalf of users with different goals and skill levels. The system will combine objective and subjective quality indices to automatically identify the manageable yet appealing subsets of document sentences to include in the summaries. To this aim, users can navigate the summary collection and provide feedback on their quality and interestingness. The collected suggestions and the summary evaluations collected during past user interactions as well as domain-specific feedback (considering differences in financial opinions based on their diverse background and specialization) can be stored and combined with objective quality indices to guide the generation of future summaries.

SNS consists of four main blocks: (i) The news crawler, which retrieves and prepares financial news to the next summarization process. (ii) The news summarizer, which selects multiple-level and multilingual summaries for the financial actors based on objective quality indices. (iii) The summary analyzer, which defines and computes ad hoc Key Performance Indicators (KPIs) on top of the generated summaries to support manual summary exploration. (iv) The collection and analysis of subjective quality indices, which collects feedback on summaries to drive the generation of future summaries according to the user’s expectations.

In this paper we present a preliminary implementation of the first three blocks. The system offline analyzes a potentially large corpora of financial news written in different languages collected through ad-hoc developed crawler. To characterize the context in which the underlying companies operate, stocks are clustered into upper-
level classes to support the generation of summaries at different abstraction levels. For example, according to
the analyst’s needs, stocks can be clustered into the corresponding financial sector (e.g. Customer goods,
Financial, Basic materials) or into the corresponding market index (e.g. Standard and Poor 500, DAX). To
the news associated with each stock under analysis, any general-purpose multilingual summarization algorithms
be integrated in SNS and applied. We adopted a divide-and-conquer strategy to deal with news in diverse
languages, i.e., we partition the set of news into homogeneous subsets and generates one summary per language,
which can be explored in place of the (potentially large) news corpora. In the current system version, analysts
read through the most salient news content written in the same language as it appears in the original news.
However, thanks to the increasing power of summarization systems, we envision an extension of the system
towards a cross-lingual environment. In the current implementation the itemset-based MWI-Sum summarization
algorithm Baralis et al. [2015] is applied to the news corpora to generate one digest per language. The digest,
which is a selection of the most interesting news sentences, consists of two parts: (i) a low-level summary, which
gives more insights into the specific stock, and (ii) an high-level summary, which provides a succinct description
of the contextual information associated with the stock. While low-level summaries are generated from the
pool of news associated with specific stocks, high-level summaries are extracted from the news ranging over
the upper-level class to which the stock belongs to. The extracted summaries are explored by investors (with
different skill levels) to understand the dynamics behind stock price movements. In support of manual summary
exploration, ad hoc Key Performance Indicators are defined and computed on top of the analyzed summaries
and news. A more thorough description of each block of the SNS system is given in the sections.

3.1 The news crawler
The news crawler retrieves collections of news about market stocks from various sources. News retrieval relies on
ad hoc Application Programming Interfaces (APIs) (e.g. the YahooFinance APIs) and on Web crawling (when
no API is available). The news provider retrieves the news ranging over a list of stocks or financial sectors and
related to a given time window specified by the analyst. The Web crawler systematically browses a subset of
Web pages from the World Wide Web and extracts useful content by parsing their HTML content. Web page
selection can be either driven by search engines or tailored to specific Web portals.

Given a set \( S \) of stocks under analysis, news crawling retrieves a set of news \( N \). Each news \( n_i \in N \) is tagged
with a set of stocks \( S_j \in S \) that are mentioned in the news and it is written in a specific language \( l_k \).

Depending on the analyst’s goal, stocks can be classified into higher-level classes. Let \( C(s_j) \) be the class of
stock \( s_j \). The class can be the main market index in which stock \( s_j \) is quoted, the country of the underlying
company, or the financial sector to which the stock belongs to. For example, stock Apple Inc. is indexed in
the Nasdaq-100 index, which groups all the main American technological shares. The underlying company is
American and its financial sector, according to YahooFinance, is Technology. For the sake of simplicity, hereafter
we will consider the financial sector given by YahooFinance as higher-level stock aggregation.

For our purposes, we cluster news ranging over the same stock \( s_j \) or class \( C_q \) and written in the same
language \( l_k \). Let \( N(s_j, l_k) \) be the set of news tagged with stock \( s_j \) and written in language \( l_k \) and let \( N(C_q,l_k) \)
be the set of news tagged with stock of class \( C_q \) and written in language \( l_k \).

3.2 The news summarizer
The step extracts a compact summary from each set of news \( N(s_j, l_k) \) and \( N(C_q, l_k) \). For each stock \( s_i \in S \)
and language \( l_k \), it generates a two-level digest of the corresponding news corpora consisting of (i) a low-level
summary, which gives more insights into the specific stock, and (ii) an high-level summary, which provides a
succinct description of the contextual information associated with the stock.

While low-level summaries are generated from the pool of news associated with specific stocks and languages
\( (N(s_j, l_k)) \), high-level summaries are extracted from the news ranging over the upper-level class to which
the stock belongs to \( (N(C_q, l_k)) \). Hence, for each stock and language the digest consists of a pair of textual summaries
describing salient information about the stock and its context, respectively.

The news summarization strategy relies on a three-step process described below:

**News preprocessing.** To suit the input news document collection to the subsequent mining process, stopword
and stemming algorithms are applied to the news corpora. Stopword elimination filters out the words having
least semantic content, because their presence would bias the quality of the next mining phase. Stemming is
an established text preprocessing technique whose aim is to reduce news words to their root form (i.e., the
stem) Tan et al. [2005]. This step, which can be enabled or disabled depending on the analyst’s preferences,
reduces the variance of the textual content to a more compact set of word roots. Stopword and stemming
algorithms are currently available for a large variety of languages.

The output of stemming and stopword elimination is modeled as a bag-of-word (BOW) Tan et al. [2005],
where each sentence of the preprocessed news corpora is represented as an unordered set of word stems.
To each stem a relevance score is assigned based on the term frequency-document frequency (tf-df) statistics Baralis et al. [2015]. The term frequency document frequency (tf-df) \( td_{zi} \) of stem \( s_{zi} \) is computed as follows:
\[
td_{zi} = \frac{o_{zi}}{|n_{i}| \cdot |\{n_{i} \in N : s_{zi} \subseteq n_{i}\}|},
\]
where \( o_{zi} \) is the number of occurrences of the \( z \)-th stem \( s_{zi} \) in the \( i \)-th news document \( n_{i} \), \( N \) is the news corpora under analysis, \( |n_{i}| \) is the number of stems that are contained in the \( i \)-th news \( n_{i} \), and \( |\{n_{i} \in N : s_{zi} \subseteq n_{i}\}| \) represents the news document frequency of the stem \( s_{zi} \) in the whole corpora.

The key idea behind this metric is that word stems that frequently occur both locally (within a news) and globally (in the corpora) get maximal td-df score, because are worth considering to summarize a set of news ranging over the same topic. Conversely, the tf-df value of a stem reduces to zero when a term never occurs in the analyzed news corpora.

**Multilingual summarization algorithm.** The summarization process consists of two steps: (i) itemset mining and (ii) sentence selection and ranking. Both steps are language-independent, provided that all the documents are written in the same language. A summary of the main algorithm steps is below. More details on the summarization algorithms are given in Baralis et al. [2015].

In our context, an itemset is an arbitrary set of word stems occurring in the news sentences. Frequent itemset mining applied to textual data allows discovering the most significant correlations between textual words Tan et al. [2005]. More specifically, frequent itemset represent combinations of stems that frequently co-occur in the analyzed data.

A model consisting of frequent weighted itemsets is generated from the preprocessed news corpora. Weighted frequent itemsets are frequent itemsets that consist of a set of word stems with averagely high importance (in terms of tf-df score). Item weights were computed according to the td-df metric Baralis et al. [2015] and used to prune the itemsets that (i) consist of lowly relevant items and (ii) rarely occur in the analyzed data.

In the second step, a subset of news sentences is selected and included in the output summary. A sentence covers an itemset if it contains the corresponding combination of word stems. Since itemsets represent the most significant underlying correlations among words, the number of covered itemsets per sentence is exploited as evaluation criterion of the sentence relevance in the news corpora.

To generate a summary consisting of the most salient news content, the sentences that cover the largest number of weighted frequent itemsets are iteratively selected until all the itemsets in the model are covered by at least one sentence. The order of appearance of the sentences in the summary reflects their relative importance, i.e., sentences covering the largest number of itemsets are placed first in the output summary.

### 3.3 The summary analyzer

The generated summaries can be exploited to understand the dynamics behind stock price movements. To support analysts in summary inspection, a set of Key Performance Indicators (KPIs) is defined on top of the output summaries and the original news corpora. KPIs can be intra-summary, if they exclusively refer to a single summary and its original news corpora, or inter-summary if they compare multiple summaries (e.g. low-level vs. high-level summaries).

Examples of intra-summary KPIs are the following ones:

1. **List of stocks/financial sectors cited in the summary and number of times each stock/sector is cited in the summary and in the original news corpora.**
2. **Top-10 most frequently occurring word stems and number of times they occur in the summary and in the original news corpora.**

KPI (1) highlights the stocks/sectors that are frequently mentioned together in the news corpora. They can be useful for recommending further interesting readings on the topic. For example, if a stock is frequently cited in most of the explored summaries, it is worth reading through the corresponding summary as well. Similarly, the price movements of the stocks of a sector can be correlated with those of another sector. KPI (2) reports most frequently occurring word stems. They are worth considering to characterize the context in which a company operate and the reasons behind unexpected market price movements.

Examples of inter-summary KPIs are the following ones:

3. **List of stocks/financial sectors cited in both the low- and high-level summaries and number of times they co-occur.**
4. **List of word stems cited both in the low- and high-level summaries and number of times they co-occur.**
5. **Top-10 most frequently occurring terms in the low-level (high-level, resp.) summary not occurring in the top-10 list of the high-level (low-level, resp.) summary.**

Since the price movements of a stock are often correlated with those of the corresponding financial sector, it is worth analyzing stock/sector co-citations. For example, word stems that frequently occur in the low-level (stock-oriented) summary and not in the high-level (sector-oriented) characterize stock price movements independently of the underlying sector. On the other hand, word stems that frequently occur at both levels are peculiar of the sector trend.
4 Case study

To validate the effectiveness and usability of the proposed approach, we conducted experiments on real financial news crawled from YahooFinance and GoogleNews.

**Experimental design.** From YahooFinance we crawled all the English-written news published from December 1st, 2015 to February 29th, 2016 related to the stocks indexed by the main American and Italian market indices, namely SP-500 and FTSE MIB-40. From GoogleNews we selected the 10 top-ranked Italian-written news published in the same period on each stock of the FTSE MIB-40 index. Stocks were classified according to the categorization provided by YahooFinance.

All the experiments were performed on a 3.0 GHz 64 bit Intel Xeon PC with 4 GB main memory running Ubuntu 10.04 LTS (kernel 2.6.32-31). To perform stopword elimination on the news documents, in our experiments we used the Natural Language Toolkit (NLTK) stopword corpus, whereas to perform stemming we adopted the Lucene stemmer McCandless et al. [2010]. To generate the summaries we used the implementation of the multilingual summarizer provided by the respective authors. When not otherwise specified, we considered the standard configuration setting indicated in Baralis et al. [2015]. For all the performed experiments, the time spent in summary generation varied between few seconds and few minutes, depending on the number of considered news and the data distribution.

**Data characterization.** As a case study, hereafter we will consider the 5 most cited SP-500 and FTSE MIB-40 stocks (according to YahooFinance) during the analyzed period. Table 1 reports the information about the analyzed stocks, the corresponding classes (i.e., the financial sectors), the number of news in which the stocks were cited, and the length of the low- and high-level summaries.

The 10 representative stocks belong to 6 different financial sectors. As expected, the cardinality of the sets of news associated with financial sectors is approximately one order of magnitude higher than those corresponding to individual stocks. Italian-written news are, on average, longer than English-written ones. Summary lengths range between 4 to 8 sentences for English-written summaries and from 6 to 15 sentences for Italian-written summaries. In some cases, the length of the high-level summaries is lower than those of the low-level ones, because a larger set of candidate sentences is available. Since the summarizer selects first the sentences that cover a largest amount of frequent weighted itemsets, when coping with similar data distributions selecting a more compact set of sentences to include in the summary is easier while coping with larger corpora.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Class</th>
<th>Per-stock # news</th>
<th>Per-class # news</th>
<th>Low-level summary length (# chars)</th>
<th>High-level summary length (# chars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Electric Company</td>
<td>Industrial Goods</td>
<td>102</td>
<td>844</td>
<td>1320</td>
<td>1120</td>
</tr>
<tr>
<td>E. I. du Pont de Nemours and Company</td>
<td>Basic Materials</td>
<td>53</td>
<td>734</td>
<td>1716</td>
<td>1163</td>
</tr>
<tr>
<td>Chevron Corporation</td>
<td>Basic Materials</td>
<td>42</td>
<td>734</td>
<td>1859</td>
<td>1163</td>
</tr>
<tr>
<td>Cisco Systems Inc.</td>
<td>Technology</td>
<td>38</td>
<td>1,525</td>
<td>2008</td>
<td>1423</td>
</tr>
<tr>
<td>Caterpillar Inc.</td>
<td>Industrial Goods</td>
<td>19</td>
<td>844</td>
<td>2404</td>
<td>1120</td>
</tr>
<tr>
<td>Fiat Chrysler Automobiles N.V.</td>
<td>Consumer Goods</td>
<td>87</td>
<td>1530</td>
<td>1921</td>
<td>1535</td>
</tr>
<tr>
<td>ENI S.p.A.</td>
<td>Basic Materials</td>
<td>19</td>
<td>844</td>
<td>3783</td>
<td>1133</td>
</tr>
<tr>
<td>Saipem S.p.A.</td>
<td>Basic Materials</td>
<td>17</td>
<td>734</td>
<td>7855</td>
<td>1133</td>
</tr>
<tr>
<td>Tenaris S.p.A.</td>
<td>Basic Materials</td>
<td>12</td>
<td>734</td>
<td>217</td>
<td>1133</td>
</tr>
<tr>
<td>Terna S.p.A.</td>
<td>Utilities</td>
<td>11</td>
<td>109</td>
<td>273</td>
<td>6645</td>
</tr>
<tr>
<td>Fiat Chrysler Automobiles N.V.</td>
<td>Consumer Goods</td>
<td>10</td>
<td>40</td>
<td>4132</td>
<td>5126</td>
</tr>
<tr>
<td>ENI S.p.A.</td>
<td>Basic Materials</td>
<td>10</td>
<td>30</td>
<td>5853</td>
<td>3637</td>
</tr>
<tr>
<td>Saipem S.p.A.</td>
<td>Basic Materials</td>
<td>10</td>
<td>30</td>
<td>4082</td>
<td>3637</td>
</tr>
<tr>
<td>Tenaris S.p.A.</td>
<td>Basic Materials</td>
<td>10</td>
<td>30</td>
<td>3879</td>
<td>3637</td>
</tr>
<tr>
<td>Terna S.p.A.</td>
<td>Utilities</td>
<td>10</td>
<td>40</td>
<td>3503</td>
<td>5944</td>
</tr>
</tbody>
</table>

**Summary analysis.** Table 2 reports two examples of digests consisting of pairs of low- and high-level English-written summaries generated from the YahooFinance news related to stocks ENI S.p.A. and Saipem S.p.A. The underlying companies all belong to sector Basic Materials and are both global leaders in drilling services, as well as in the engineering, construction, and installation of pipelines in the oil and gas market.

Low-level summaries contain good news for the underlying companies. For example, the announces of the commercial agreements of ENI S.p.A. and Saipem S.p.A. with Iran are positive signals, which may foster medium- and long-term investments from traders, funds, and bankers.

The analysis of the top-ranked sentences of the high-level summary can be useful for competitive intelligence. For example, for Italian oil and gas market companies knowing that U.S. producers (e.g. Noble Energy Inc) are achieving unexpectedly good profits is important to plan future investments. This kind of information cannot be inferred from single-stock news corpora and their low-level summaries, unless perusing the summaries associated with each share belonging to that sector.

KPI-based text analyses can provide insightful information about summary content. For example, Table 3...
results, achieved on real multilingual news corpora, demonstrate the effectiveness of the proposed approach in different abstraction levels to highlight knowledge fruitful for different financial actors and scopes. Preliminary reports, for three representative financial sectors (among those occurring in Table 1), the most frequently cited stocks appearing in the corresponding high-level summary. Specifically, from the set of stocks appearing the summary, the top-5 stocks in order of decreasing number of citing news (i.e., the number of news, belonging to that class corpora, in which the stock is mentioned at least once) are reported. For each cited stock, the corresponding financial sector and the total number of occurrences in the news corpora are also given.

For the most targeted sectors Technology and Basic Materials the majority of the top ranked cited stocks belong to the corresponding sector (e.g. for Technology stocks Google Inc., TENCENT, and Infineon Technologies AG), whereas for the more general sector Industrial Goods most citations link to stocks belonging to other (pertinent) classes. Citations in the high-level summary may prompt further readings and deepening. For example, based on the citations in class Technology, the low-level summaries of stocks Google Inc., Sealed Air Corporation, and Ebay Inc. are worth reading through. Note that since some of the mostly cited stocks do not belong to class Technology, lower-level summaries are likely to contain complementary content with respect to the high-level summary of class Technology. Cross-citations between different financial sectors can be also helpful for driving portfolio diversification strategies, i.e., to invest on companies related to different markets or sectors in order to spread bets across a larger number of financial assets. For example, if oil and gas market stocks are potentially highly profitable in the long term, stocks cited in the corresponding high-level summary but not belonging to class Basic Materials (e.g. Sealed Air Corporation) can be considered to build a diversified stock portfolio.

### Table 2: Multiple-level summary examples.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Summary (Italian companies will sign commercial agreements with Iran worth between 15 billion and 17 billion euros during President Hassan Rouhani’s visit in Rome this week, an Italian government source said on Monday.)</th>
<th>Summary (Saipem SpA is nearing a plan to sell new shares worth as much as 3.5 billion euros as the Italian oil-services company seeks to shore up its balance sheet amid a slump in crude prices, according to people familiar with the matter.)</th>
<th>Summary (Among the deals struck on Monday were a pipeline contract worth between $4 billion and $5 billion for oil services group Saipem, up to 5.7 billion euros in contracts for Italian steel firm Danieli and up to 4 billion euros of business for firm Condotte d’Acqua.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Any additional Libyan output would feed a glut that has seen oil prices slump by more than 65 percent since June last year, and even the prospect of production coming back could spark further declines, Mallinson said.</td>
<td>The global oil industry is set to repeat this year’s $200 billion of investment cuts in 2016, raising even more concerns than the current slump in crude prices, according to the chief executive officer of Italy’s Eni SpA.</td>
<td>The company, which bought Rosetta Resources in a $2 billion deal last year, said sales volumes rose 8 percent to 422,000 barrels of oil equivalent per day, pro-forma for the deal.</td>
</tr>
<tr>
<td>2nd</td>
<td>Italian companies will sign commercial agreements with Iran worth between 15 billion and 17 billion euros during President Hassan Rouhani’s visit in Rome this week, an Italian government source said on Monday.</td>
<td>Oil and gas producer Noble Energy Inc posted a surprise quarterly profit and said it would “monetize” assets to cope with a slump in oil prices that is eroding cash flows, sending the company’s shares up 8 percent.</td>
<td>As of earlier this month, SandTilde’s shares are no longer listed on the New York Stock Exchange, and trade on the OTC Pink marketplace instead with a market capitalization of around $30 million.</td>
</tr>
<tr>
<td>3rd</td>
<td>In this paper we address the problem of summarizing large collections of textual news on stock markets written in different languages. The preliminary version of the proposed engine performs multilingual summaries at different abstraction levels to highlight knowledge fruitful for different financial actors and scopes. Preliminary results, achieved on real multilingual news corpora, demonstrate the effectiveness of the proposed approach in</td>
<td>Top-3 sentences in the high-level summary of class Basic Materials</td>
<td>Top-3 sentences in the high-level summary of class Technology</td>
</tr>
</tbody>
</table>

### Table 3: Per-class top-5 cited stocks.

<table>
<thead>
<tr>
<th>Stock (YahooFinance ID)</th>
<th>Class</th>
<th>Num. of occurrences in the news</th>
<th>Num. of citing news</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Inc. (GOOGL)</td>
<td>Technology</td>
<td>4049</td>
<td>1008 out of 1525</td>
</tr>
<tr>
<td>Sealed Air Corporation (SEE)</td>
<td>Consumer Goods</td>
<td>945</td>
<td>568 out of 1525</td>
</tr>
<tr>
<td>eBay Inc. (EBAY)</td>
<td>Services</td>
<td>34</td>
<td>18 out of 1525</td>
</tr>
<tr>
<td>TENCENT (0700.HK)</td>
<td>Technology</td>
<td>40</td>
<td>18 out of 1525</td>
</tr>
<tr>
<td>Infineon Technologies AG (IFX.DE)</td>
<td>Technology</td>
<td>9</td>
<td>5 out of 1525</td>
</tr>
<tr>
<td>AGL Resources Inc. (GAS)</td>
<td>Utilities</td>
<td>748</td>
<td>252 out of 734</td>
</tr>
<tr>
<td>Nucor Corporation (NUE)</td>
<td>Basic Materials</td>
<td>449</td>
<td>240 out of 734</td>
</tr>
<tr>
<td>ENI S.p.A. (ENI.MI)</td>
<td>Basic Materials</td>
<td>247</td>
<td>173 out of 734</td>
</tr>
<tr>
<td>Noble Energy Inc. (NBL)</td>
<td>Basic Materials</td>
<td>58</td>
<td>27 out of 734</td>
</tr>
<tr>
<td>Reynolds American Inc. (RAI)</td>
<td>Utilities</td>
<td>408</td>
<td>225 out of 844</td>
</tr>
<tr>
<td>Apache Corporation (APAC)</td>
<td>Basic Materials</td>
<td>409</td>
<td>209 out of 844</td>
</tr>
<tr>
<td>National Oilwell Varco, Inc. (NOV)</td>
<td>Basic Materials</td>
<td>191</td>
<td>132 out of 844</td>
</tr>
<tr>
<td>Nike, Inc. (NKE)</td>
<td>Consumer Goods</td>
<td>65</td>
<td>47 out of 844</td>
</tr>
<tr>
<td>SAP SE (SAP.DE)</td>
<td>Technology</td>
<td>62</td>
<td>43 out of 844</td>
</tr>
</tbody>
</table>

### 5 Conclusions and future works

In this paper we address the problem of summarizing large collections of textual news on stock markets written in different languages. The preliminary version of the proposed engine performs multilingual summaries at different abstraction levels to highlight knowledge fruitful for different financial actors and scopes. Preliminary results, achieved on real multilingual news corpora, demonstrate the effectiveness of the proposed approach in
generating interesting summaries, which may enhance the awareness of private and public investors on foreign markets. Currently, we are planning to extend the current version of the architecture towards a cross-lingual environment, which will produce a unified summary of all the multilingual news collections written in the native language of the final user. An adaptive sentence selection and ranking algorithm will be also developed and integrated to choose the most interesting content based on both objective and subjective quality indices.

References


