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 Filippova et al. [2009], Otterbacher et al. [2006], Yang and Wang [2003]. To facilitate short-term trading activities, in Filippova 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 Filippova 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

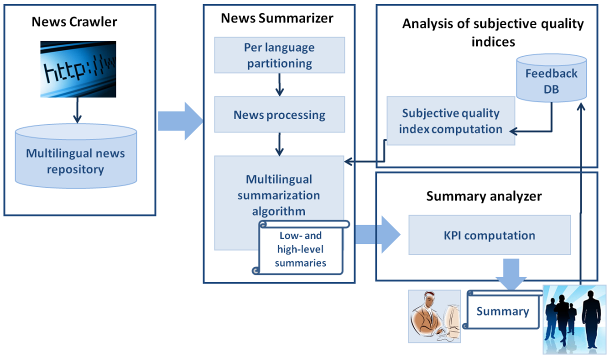
News Crawler
http://w
Multilingual news repository
News Summarizer
Per language partitioning
News processing
Multilingual summarization algorithm
Low- and high-level summaries
Analysis of subjective quality indices
Feedback DB
Subjective quality index computation
Summary analyzer
KPI computation
Summary

Figure 1: The Stock News Summarizer architecture.

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 can 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 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_i 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: $td_{zi} = \frac{o_{zi}}{|n_i|} \cdot \frac{|\{n_i \in N \,:\, s_{zi} \in n_i\}|}{|N|}$, 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 $\frac{|\{n_i \in N \,:\, s_{zi} \in n_i\}|}{|N|}$ 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 1: Analyzed stocks and news.

Stock	Class	Per-stock # news	Per-class # news	Low-level summary length (# chars)	High-level summary length (# chars)
English-written news					
SP-500 stocks					
General Electric Company	Industrial Goods	132	844	1420	1120
E. I. du Pont de Nemours and Company	Basic Materials	53	734	1716	1163
Chevron Corporation	Basic Materials	42	734	1859	1163
Cisco Systems Inc.	Technology	38	1,525	2008	1423
Caterpillar Inc.	Industrial Goods	19	844	2404	1120
FTSE MIB-40 stocks					
Fiat Chrysler Automobiles N.V.	Consumer Goods	87	1530	1521	1535
ENI S.p.A.	Basic Materials	19	844	3783	1133
Saipem S.p.A.	Basic Materials	17	734	7855	1133
Tenaris S.p.A.	Basic Materials	12	734	217	1133
Terna S.p.A.	Utilities	11	109	273	6645
Italian-written news					
FTSE MIB-40 stocks					
Fiat Chrysler Automobiles N.V.	Consumer Goods	10	40	4132	5126
ENI S.p.A.	Basic Materials	10	30	5853	3637
Saipem S.p.A.	Basic Materials	10	30	4082	3637
Tenaris S.p.A.	Basic Materials	10	30	3879	3637
Terna S.p.A.	Utilities	10	40	3503	5944

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

Table 2: Multiple-level summary examples.

Top-3 sentences in the low-level summary of Saipem S.p.A.	
1st	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.
2nd	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.
3rd	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.
Top-3 sentences in the low-level summary of ENI S.p.A.	
1st	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.
2nd	Italian companies will sign commercial agreements with Iran worth between 15 billion and 17 billion euros ($16.2 and $18.4 billion) during President Hassan Rouhani's visit in Rome this week, an Italian government source said on Monday.
3rd	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.
Top-3 sentences in the high-level summary of class Basic Materials	
1st	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.
2nd	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.
3rd	As of earlier this month, SandRidge'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.

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 3: Per-class top-5 cited stocks.

Stock (YahooFinance ID)	**Class**	**Num. of occurrences in the news**	**Num. of citing news**
class Technology			
Google Inc. (GOOGL)	Technology	4049	1008 out of 1525
Sealed Air Corporation (SEE)	Consumer Goods	945	568 out of 1525
eBay Inc. (EBAY)	Services	34	18 out of 1525
TENCENT (0700.HK)	Technology	40	18 out of 1525
Infineon Technologies AG (IFX.DE)	Technology	9	5 out of 1525
class Basic Materials			
Sealed Air Corporation (SEE)	Consumer Goods	696	311 out of 734
AGL Resources Inc. (GAS)	Utilities	748	252 out of 734
Nucor Corporation (NUE)	Basic Materials	449	240 out of 734
ENI S.p.A. (ENI.MI)	Basic Materials	247	173 out of 734
Noble Energy Inc. (NBL)	Basic Materials	58	27 out of 734
class Industrial Goods			
Reynolds American Inc. (RAI)	Utilities	408	225 out of 844
Apache Corporation (APA)	Basic Materials	409	206 out of 844
National Oilwell Varco, Inc. (NOV)	Basic Materials	191	132 out of 844
Nike, Inc. (NKE)	Consumer Goods	65	47 out of 844
SAP SE (SAP.DE)	Technology	62	43 out of 844

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

Elena Baralis, Luca Cagliero, Saima Jabeen, and Alessandro Fiori. Multi-document summarization exploiting frequent itemsets. In *Proceedings of the ACM Symposium on Applied Computing, SAC 2012, Riva, Trento, Italy, March 26-30, 2012*, pages 782–786, 2012. doi: 10.1145/2245276.2245427. URL http://doi.acm.org/10.1145/2245276.2245427.

Elena Baralis, Luca Cagliero, Alessandro Fiori, and Paolo Garza. Mwi-sum: A multilingual summarizer based on frequent weighted itemsets. *ACM Trans. Inf. Syst.*, 34(1):5, 2015. doi: 10.1145/2809786. URL http://doi.acm.org/10.1145/2809786.

Document Understanding Conference. HTL/NAACL workshop on text summarization, 2004.

Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Company-oriented extractive summarization of financial news. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 246–254, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609067.1609094.

Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 71–78, 2003.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0.* Manning Publications Co., Greenwich, CT, USA, 2010. ISBN 1933988177, 9781933988177.

Jahna Otterbacher, Dragomir Radev, and Omer Kareem. News to go: Hierarchical text summarization for mobile devices. In *29th ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 589–596, 2006. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148271. URL http://doi.acm.org/10.1145/1148170.1148271.

Josef Steinberger, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi, and Vanni Zavarella. JRC's participation at TAC 2011: Guided and multilingual summarization tasks. In *TAC'11: Proceedings of the The 2011 Text Analysis Conference*, 2011.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Intoduction to Data Mining.* Addison Wesley, 2005.

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. Integrating document clustering and multidocument summarization. *ACM Trans. Knowl. Discov. Data*, 5:14:1–14:26, August 2011. ISSN 1556-4681. doi: http://doi.acm.org/10.1145/1993077.1993078. URL http://doi.acm.org/10.1145/1993077.1993078.

Christopher C. C. Yang and Fu Lee Wang. Automatic summarization of financial news delivery on mobile devices. In *ACM Conference on the World Wide Web*, pages 225–233, 2003.