Multi-Entity Polarity Analysis and Detection of Subjectivity in Financial Documents

Josiane Rodrigues[1], Marco Cristo[1], Javier Zambrano Ferreira[2]
David Fernandes[1], André Carvalho[1]

[1] Universidade Federal do Amazonas, Brasil
{josiane, marco.cristo, david, andre}@icomp.ufam.edu.br
[2] Instituto Ambiental e Tecnológico da Amazônia, Brasil
javier.ferreira@iatecam.org.br

Abstract. Polarity analysis aims at classifying an author's opinion as positive, negative, or neutral. However, given the sheer volume of information available on the web, manually carrying out such task is unfeasible. In particular, this type of analysis is useful for companies when making decisions related to the financial market, which is particularly prone to changes according to shifting moods and opinions. Most studies in the literature deal with this problem by considering that whole documents have a single, global polarity. However it is not unusual that documents have opinions on several entities with possibly different polarities. This suggests that the polarity classification should be performed in an entity level. We also noted that many financial documents may not emit any opinion. Therefore, in this paper we propose a supervised polarity classification method based on multiple models and detection of subjectivity, in order to deal with financial documents that cite multiple entities. Our results showed that the hierarchical, multiple-models approach significantly outperformed the global-model baseline.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: Miscellaneous

Keywords: Anaphora Resolution, Detection of Subjectivity, Machine Learning, Sentiment Analysis, Web Data Annotation

1. INTRODUCTION

The internet is a large repository of mostly unstructured textual information, and in order to take advantage of this information it is important to have a deeper understanding of its content. However, its lack of structure, allied with the scale of its data and its rate of growth, makes any effort on manually analyzing its content unfeasible. Thus arose the need of deploying automatic techniques to analyze textual content, among which polarity analysis has received growing interest from the research community and the industry [Becker and Tumitan 2013].

Polarity analysis consists in determining the polarity of an opinion's author about the subject under discussion, such as inferring whether the opinion is favorable, neutral, or unfavorable about that subject. Opinions usually show what the author actually thinks about given a subject, and this information can be applied in a myriad of domains, such as the financial, online shopping and even the musical domains [Liu 2012]. For instance, it might be useful to automatically infer (a) the opinion of a reviewer about a movie from his review, (b) the opinion of a client about a certain product from the reviews that he posted on an online store, and (c) the opinion of a person about something that was posted on a social network.

THE ANALYSIS: The verdict against Samsung could give Nokia an edge. The Windows Phone is substantially different from Apple's iPhone operating system and hasn't landed in its legal sights, and some Wall Street analysts say that the verdict against Samsung is likely to slow growth of smartphones that run on Android.

Fig. 1. Financial text excerpt citing multiple entities. Content by Associated Press[1].

In this work, we will focus on the polarity analysis of documents in the financial domain. In financial documents, it is common that opinions are issued about the current and future situation of companies (entities of interest). These opinions are important, since they might reflect the upcoming performance of those entities in stock markets. Usually, a single financial document may cite a number of entities, which implies that a document can present as many polarities as the number of entities it cites. Another specific aspect of financial documents, in contrast with the purely editorial nature of documents such as book and movie reviews, is that they do not always express polar opinions about entities, often expressing neutral opinions. In general, these opinions have less impact on the sentiment about of the entities that they cite. Thus, a first task of interest in this area is to identify the documents on which opinions are expressed. Once those documents are identified, a second task would be to classify their polarities.

In the excerpt shown in Figure 1, three entities are cited: *Samsung*, *Nokia* and *Apple*. Observe that the polarity of each entity cited is different, since it reports that Nokia can profit with the loss of Samsung in a trial involving Apple and Samsung. In particular, this document is neutral to Apple (since it is only mentioned as the iPhone manufacturer), while it is positive regarding Nokia and negative regarding Samsung. However, it is common in previous works to infer the polarity of the text as a whole regarding only a predetermined entity (e.g., specified using a query). In many of these works, a supervised method is applied to learn a model that describes the polarity of any entity, and is given examples of previously classified documents (where it is common that the label assigned to the document indicates its global polarity regarding a single entity). Thus, a single global model of polarity is learned for the collection. However, this approach does not work well when we have documents similar to the example above, in which the polarities are distinct for each entity.

In order to solve this problem, a multiple-model strategy to describe polarities could be considered, with one model for each entity. The model of a certain entity could be created from documents (or fragments of documents) in which this entity is cited. The labels of training examples should be labels defined to the entities instead of labels defined for documents. As part of the classification process, a hierarchical approach could be adopted to separate only documents that express opinions. The goal of this approach is to improve the polarity classifier learning to separate documents that issue no opinion regarding the entities.

Based on these ideas, in this work, we propose and evaluate a multiple-model learning method for the opinion detection task and polarity classification in financial documents that cite multiple entities. In particular, our contributions are: (i) a supervised hierarchical method to infer polarity at the entity level, (ii) the use of segmentation techniques in text document in order to isolate segments referring to only target entities and (iii) the use of a state-of-art unsupervised algorithm for polarity classification.

This article is an extension of our previous work published in [Ferreira et al. 2014], where we proposed a method to infer entity polarity on multi-entity documents. We, expanded this idea, proposing supervised *hierarchical* method to infer polarity at the entity level, also including experiments including the unsupervised method proposed by Moilanen and Pulman [2009], which is used as an additional baseline. Finally, we have fixed errors found in the first study concerning the experiments. In that previous work, two experiments were performed, dividing the experimental collection of documents in two sets: one considering the entire dataset and one considering only 617 documents that cite at

[1]Associated Press: http://www.ap.org/

least two entities. However, in a later experiments we verify that, for 4 of the 5 entities considered in the study, all documents should be considered as multi-entity, having the same experimental results in both sets. Thus, in the experiments that we present in Section 5 we do not make this separation anymore, considering that the entire dataset is composed of multi-entity documents.

The remainder of this article is organized as follows. In Section 2, we discuss previous works related to the topics covered in our study. In Section 3, we provide background information on polarity analysis and anaphora resolution. In Section 4, we describe the proposed solution to the problem presented. The experiments carried out and their results are reported in Section 5. Finally, in Section 6, we present our conclusions and future works.

2. RELATED WORK

As far as we know, the first study about polarity analysis in the financial domain was proposed by Azar [2009], motivated by the possibility of forecasting reactions of the stock market. By using Natural Language Processing (NLP) heuristics along with SVM classifiers and Decision Trees, the model achieved a performance comparable to human annotators. In the conducted study, they only considered documents that cited companies with more than 20 news articles in the *Reuters Key Developments Corpus*. They also observed that the models learned for the financial domain achieve lower performance levels in other domains.

Bollen et al. [2011] studied the correlation between the opinion polarity of mentions of a company and its performance in the stock market. The information about the companies was obtained from Twitter and they deployed a simple strategy based on Google Profile of Mood States (GPOMS). After analyzing a large amount of data from Twitter, they found out that the emotional changes detected in tweets correlate to changes observed in the stock market. Many other works after that have also been focusing on Twitter, such as the study performed by Montejo-Ráez et al. [2014], that used a lexical approach based on Wordnet, and a comparison of methods presented in [Araújo et al. 2013]. Schumaker and Chen [2009] also studied the problem of polarity classification in the financial domain. The authors studied two textual representations: bag-of-words and noun phrases with entities names. In their work, the authors also observed a correlation between the future prices of a company stocks and the polarity of news citing them. Another bag-of-words approach in the financial domain was proposed by Im et al. [2013], where the focus was the use of stemming, and the authors observed that it leads to improvements in sentiment recognition.

The studies presented above treated polarity as a concept related to the general opinion expressed in documents. While this is common in editorial documents where the opinions are related to a central topic, product or service, this is not true for non-structured documents that might express opinions about any number of entities.

One approach that is common among the few works that explore the polarity analysis in documents with multiple entities is the use of compositional strategies. In those strategies, the polarity is estimated within the sub-contexts and the global polarity is obtained by a composition of these sub-contextual polarities, using a dependency grammar. It should be noted that once each atomic context is associated with the polarity of an entity, this method naturally infers polarities for several entities. In this approach, the document to be classified is first broken into sentences. Each sentence is associated with a logical polarity between three possible values (positive, negative and neutral) according to a sentiment grammar. The sentence polarities are then combined on a pairwise basis, according to a dependency grammar, until a final polarity is obtained. The process starts with a lexical analysis at the word level, continues recursively on intermediary syntactic levels and ends at sentence level. This strategy was proposed for the first time by Moilanen and Pulman [2009]. After evaluation, the authors concluded that the proposed method presents an accuracy slightly smaller than that of human annotators.

Later on [Gryc and Moilanen 2010], the same authors applied their techniques to infer the polarity in the political domain on blog messages. A similar approach was proposed by Romanyshyn [2013], where a rule system was used to detect sentiments of individual clauses in reviews in the Ukrainian language. The composition of sentiment clauses allowed for a multi-entity analysis. Finally, Ward et al. [2011] proposed a framework for sentiment analysis at the entity level. The main rule of the analysis is that a specific entity is used as query for a data set. The entity sentiment is obtained from the information set in the data set returned with previously analyzed answers.

In our work we also propose a hierarchical classification approach for the task of polarity analysis. We were motivated by the fact that this task can be seen as two distinct classifications: (i) detection of subjective cases and (ii) classification of subjective cases into positive or negative. In this kind of approach, the learning algorithm considers the hierarchical relationships between the classes. A similar approach is used in the work of Pang and Lee [2004], where the polarity classification is improved by the removal of objective sentences from the training set. In that study, the authors are interested in classifying the polarity of movie reviews. To do that, they first applied a subjectivity detector that determine if each sentence is subjective or not, discarding the objective ones and creating a summary that should better represent the content of the movie reviews, which improves the performance of the used polarity classifier.

For the subjectivity classification for the Arabic language, Abdul-Mageed et al. [2011] performed binary classification at the sentence level. They used a manually annotated dataset and built a lexicon for the Arabic language composed by 3982 adjectives, which was elaborated from news articles (extracted from Penn Arabic tree bank). In a later work, Abdul-Mageed et al. [2012] extended their work to social media contents, including chat sessions, tweets, Wikipedia discussions and online forums.

Other works also explore the subjectivity concept. In [Mourad and Darwish 2013] the authors proposed an approach for polarity classification using detection of subjectivity in microblogs and Refaee and Rieser [2014] explored the same concepts for Twitter. In our work, distinctly from those works, we found that while the subjectivity detector in the first phase is very effective in determining the neutral cases, its error rate when classifying polar cases is still high enough to prevent it from having the same effectiveness of the polarity binary classifier in the second phase. Thus, a semi-hierarchical method seems more suitable.

3. BACKGROUND

3.1 Polarity Analysis

According to Liu [2012], polarity analysis aims to determine the attitude of an agent (opinion owner) regarding a specific target (opinion receiver), which can be, for instance, a topic, a product, a service, an entity or even a property of those objects. The target is usually a textual content of a specific context and time. Although the attitude of an agent can be very nuanced and complex, it is common to model it as a classification task with symbolic values such as positive, negative and neutral. From the operational point of view, polarity analysis implies the use of textual analysis techniques, natural language processing and computational linguistics to identify, extract and understand the subjectivity of the content. In the literature, the polarity analysis task is addressed in two major ways: lexical techniques, based on dictionaries, and techniques based on machine learning.

In lexical approaches, the polarity is inferred from words. In this kind of approach, the classification phase is based on the use of lexicons (dictionaries) of sentiments, compilations of words or expressions of sentiments associated with their respective polarities. In general, these techniques are unsupervised and dependent on preprocessing operations and specific transformations, such as n-gram recognition, extraction of features, irrelevant terms elimination, transformation of text in term vectors etc.

Besides lexical techniques, a number of approaches are based on modeling the problem as a machine learning task. In them, the problem of polarity analysis can be seen as a supervised classification task and many sources of information used by previous methods are used as features to represent the documents to be classified. The classification process is divided in two main steps: (i) learn a classification model from a training collection where polarity labels were previously identified (the ground truth) and (ii) predict the polarities for each entity in new documents with the resulting model from (i).

As in [Azar 2009], we adopted Support Vector Machine (SVM) [Baeza-Yates and Ribeiro-Neto 2011; Witten et al. 2011] as our polarity classifier since it has been commonly used in that task. To classify documents with SVM, we use the tool LIBSVM[2] [Chang and Lin 2011].

3.2 Anaphora Resolution

One simple strategy to determine if an entity is cited in a sentence is through verifying the occurrence of the string corresponding to the entity name in the n-grams extracted from the sentence. For example, consider the sentence "Cats are very clean" and "However, they always get dirty when they go outside". It is possible to infer that the first sentence refers to "Cats" by the simple verification of the occurrence of the string "Cats". However, this approach doesn't work for the second sentence, where the entity "Cats" was represented by the pronoun "they". The problem of determining that various texts fragments refer to the same entity is called Anaphora Resolution or Co-reference Resolution. This problem happens due to the fact that one single entity can be referred by different linguistics expressions. In the previous sentences, "Cats" and "they" are related to the entity "Cats".

Many studies in natural language processing [Poesio et al. 2011; Ng 2010] have addressed the anaphora resolution problem. More formally, the anaphora problem can be defined as: a noun A is an anaphoric antecedent of B if and only if A is needed for the interpretation of B. To solve the anaphora, the text is first segmented into sentences using a sentence splitter. Then, the elements of the sentence (such as entity names, nouns, verbs and adverbs) are identified. The identification of the entity names and nouns is essential to solve the anaphora, since they are commonly used to describe people, places, objects and concepts. This identification requires knowledge about the grammar of the target language, which describes usage patterns and specific domain information. For example, in English, the nouns can be identified through the recognition of other linguistic structures, such as definite pronouns (e.g. "Ross bought {a MP3 player / three flowers} and gave {it / them} to Nadia for her birthday}"), indefinite pronouns (e.g. "one" in "Kim bought a t-shirt so Robin decided to buy one as well"), demonstrative pronouns (e.g. "that"), nominal pronouns (e.g. "a man"", "a woman" and "the man") and proper name (e.g. "John", "Mary").

In this work, we explored Anaphora Resolution using the Stanford CoreNLP[3] tool [Manning et al. 2014]. CoreNLP provides a set of tools for natural language analysis, that integrates most of the steps of natural language processing, including: part-of-speech (POS) tagger, named entity recognizer (NER), natural language analyzer, that analysis the grammatical structure of sentences, co-reference resolution system, sentiment analysis and a bootstrapped pattern learning. The basic distribution of the tool provides support to English text analysis. However, it is possible to apply the tool in other languages, such as Chinese and Spanish. The architecture of CoreNLP system includes all the required modules to solve co-references. The processing of a text is made by following the steps:

—The input text undergo an annotation process, that consists of a sequence of annotators.
—In this annotation process, the text is represented by a sequence of tokens, that is then grouped into sentences. The tokens are then labeled with their parts of speech. Then, the tool generates

[2]http://www.csie.ntu.edu.tw/cjlin/libsvm/
[3]http://nlp.stanford.edu/software/corenlp.shtml#About

lemmas, recognizes the entities (companies names, person, places, etc.) and provides a complete syntactic analysis, including a representation of dependences based on probabilistic analysis. Based on this information, it is possible to make sentiment analysis applying a compositional model based on a classifier and implement the detection of mentions and co-reference resolutions.

—The output is an annotation containing all the information analyzed by the annotators, structured in a XML file.

4. MULTI-ENTITY POLARITY MODEL

In this section, we present the method based on multiple learning models and detection of subjectivity.

4.1 Multiple Polarity Learning with Multiple Learning Models

We adopted the polarity model described in [Ferreira et al. 2014] as a basis for the proposed model. In it, polarity analysis is seen as a classification task where, given a document collection $\mathcal{D}$ citing m target entities E, the objective is to find m classifier functions $f_j : \mathcal{D} \implies \{+, -, N\}$, where $f_j(d_i) = y_{ij}$, $1 \leq j \leq m$. In other words, a classifier for each entity E_j. Note that documents where entity E_j does not occur probably do not contribute to the learning of function f_j, since they could hardly be considered as examples of positive, negative or neutral opinions about E_j. Thus, function f_j would be better represented by $f_j : \mathcal{D}_j \implies \{+, -, N\}$, where D_j is the subset of documents of $\mathcal{D}$ that cite entity E_j.

In a similar fashion, sentences from documents that cite a set of entities $\mathcal{E}$ should not be used as training examples to the entity E_j, $E_j \notin \mathcal{E}$. Suppose that a sentence (or a whole document) s is evaluated as negative. It would not be appropriate to use s as a negative example about E_j if s does not cite E_j. Thus, let d_i^j be a document composed of all sentences of d_i which cite E_j. We denote as $\mathcal{D}^{(j)}$ the set of all documents d_i^j, i.e., the set of all documents that are composed only by sentences which cite E_j. Given such definitions, we can rewrite f_j as $f_j : \mathcal{D}^{(j)} \implies \{+, -, N\}$.

Based on these ideas, we can now define two strategies to learn polarities of multiple entities using multiple models:

—Document based Model (DbM): where each function f_j is associated with the document collection $\mathcal{D}_j$, i.e., the set of *documents* which cite E_j.

—Sentence-Set based Model (SSM): where each function f_j is associated with the document collection $\mathcal{D}^{(j)}$, i.e., the set of documents composed only by the sentences which cite E_j. In order to discover and remove sentences that do not cite E_j, we defined sentence discarding strategies, presented in the next section.

4.1.1 *Mapping Sentences and Entities.* For the second approach (SSM), it is necessary to define which sentences cite which entity, since some sentences might not cite the entities explicitly. Based on those observations, six variants of Sentence-Set based Models have been proposed: three of them based on string occurrence heuristics and three based on Anaphora Resolution. The six variants are described next:

—SSM1: the sentence s is assigned to the entities whose names occur in s. If no entity is present in s, s is assigned to all the entities.

—SSM2: the sentence s is assigned to the entities whose names occur in s. If no entity is present in s, s is discarded;

—SSM3: the sentence s is assigned to the last cited entity if no entity is present in s. The intuition behind this heuristic is that if no new citation is done, the next sentence probably still refers to the last cited entity;

—SSM4: the sentence s is assigned to the entities directly referenced by s. If no entity is referenced by s, s is assigned to all the entities. This heuristic is equivalent to SSM1, but using anaphora resolution;
—SSM5: the sentence is assigned to all the entities referenced by s. If no entity is referenced by s, s is discarded. This heuristic is equivalent to SSM2, but using anaphora resolution.
—SSM6: the sentence s is assigned to the last referenced entity if no entity is referenced by s. This heuristic is equivalent to SSM3, but using anaphora resolution.

In strategies SSM1 to SSM3, the notion of citation corresponds to the explicit occurrence of the entity name in the sentence, verified by simple name string matching. Regarding strategies SSM4 to SSM6, we consider that a sentence cites an entity when it refers to it directly or indirectly. For those indirect citations, we applied CoreNLP to solve the anaphoras found in the sentences.

4.2 Detecting Subjectivity

In sentiment analysis, we are usually interested in the author's opinion regarding an object. Thus, a way to improve polarity classification is by removing from the training collection documents that do not issue an opinion on any object. Other works provide methods for polarity analysis, specifying whether or not a document has opinion [Pang and Lee 2004; Abdul-Mageed et al. 2011; Abdul-Mageed et al. 2012], but all those methods are focused on a whole document level analysis, with none of them being focused at the entity level. Therefore, in this work we propose a number of methods that combine the segmentation strategies previously described as a subjectivity (opinion) detection model, which is applied in a stage prior to the polarity classification. After isolating the sentences, we then can solve the problem of classifying the opinion this sentence has about a given entity as follows. Given a document d and an entity E, two tasks are performed:

—Subjective classification: determine if d is a subjective or an objective document regarding E.
—Polarity classification: if d is subjective regarding E, determine if d expresses a positive or negative opinion.

Figure 2 illustrates this process. The previously segmented documents are submitted to a classifier that will try to learn which documents are subjective, and these, in turn, are classified by its polarity. In this sense, the two classification sub-tasks are important because they (1) filter out the documents that have no opinion in relation to the entity in question; (2) once these documents are filtered out, it becomes easier to determine if the opinion expressed by them and its characteristics are positive or negative. This can prevent the polarity classifier from considering irrelevant or even deceptive

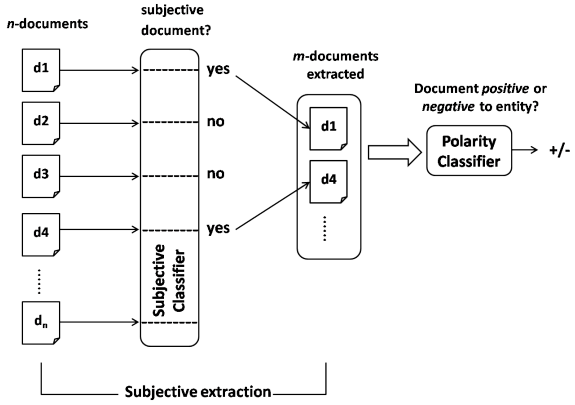

Fig. 2. Polarity detection *after* opinion detection.

Table I. Polarity distribution by page according to the ground truth.

Entity	Positive	Negative	Neutral	Total
Apple	261	131	562	954
Google	105	39	276	420
Samsung	81	58	197	337
Microsoft	55	31	195	281
Nokia	29	25	71	125
Totals	531	284	1301	2117

texts: for example, the sentence "Samsung made a deal with Apple" does not provide any clue on the author's opinion about Samsung or Apple and it should not be incorporated as positive or negative examples about the companies.

Finally, note that unlike other works in the literature, we will adopt a more flexible approach regarding the second stage of classification, considering the possibility of using different classification approaches according to the error rate observed in the first stage. For instance, instead of classifying only positive and negative polarities in the second stage, we also classify neutral polarity, since there are still documents with neutral polarity in the second stage due to the classifier errors in the first stage.

5. EXPERIMENTAL EVALUATION

In this section we evaluate the method proposed in Section 4. We present a collection over which we have done experiments and shown the criteria for choosing the baseline. Then we expose and discuss the results obtained from the comparison between the baseline and the proposed solution for the problem.

5.1 Experimental Methodology

To evaluate our method, we used a dataset comprised of one thousand documents. These documents were extracted from the following business websites and financial market news in the year 2012: Reuters[4], Bloomberg[5], Financial Times[6], Forbes[7], The New York Times[8], AllThingsD[9] and CNN Money[10]. We defined a set of target entities for the study (Apple, Google, Samsung, Microsoft and Nokia) and only pages that cite at least one of those entities were selected.

From the set of pages containing the target entities, a random subset of 1,000 pages was selected and labeled by 40 human annotators. Each annotator received 25 pages and evaluated them according to their global polarity and the polarity of each target entity present in the page. From a total of 1,000 documents, 300 were evaluated as positive, 85 as negative and 615 as neutral. Table I presents the distribution of polarities by entity and the total number of pages associated with each entity.

Note that in this dataset, as shown by table I, there is a distinct bias in both polarities (61% are neutral and only 13% are negative) and entity (Apple was cited in 95% of the documents, while Nokia in only 12%). Table II(b) presents the entities distribution by page, where each entity is a company listed in Forbes Fortune List 2012[11]. As we can observe, the majority of the documents (617, which corresponds to 62%) cite more than one entity. However, the most common number of entities per

[4]http://www.reuters.com
[5]http://www.bloomberg.com
[6]http://www.ft.com
[7]http://www.forbes.com
[8]http://www.nytimes.com
[9]http://allthingsd.com
[10]http://money.cnn.com/
[11]http://www.forbes.com/global2000/list/

Table II. Entity and polarity distribution by pages according to the ground truth.

Polarities	Pages
1	697
2	257
3	46

(a) Entity distribution

Entities	Pages
1	383
2	308
3	164
4	87
5	58

(b) Polarity distribution

document is just one (38% of the documents). Table II(a) shows the polarity distribution by page. Although the majority of the documents cite more than one entity, only in 30% of them different polarities are observed.

The manually labeled data collection was then processed according to the segmentation strategies described in Section 4.1.1. The text was segmented into sentences producing data sets corresponding to SSM1, SSM2 and SSM3 strategies. Similarly, we used CoreNLP to solve anaphora and normalize the citations related to each entity. The documents were then segmented into sentences to create the data sets corresponding to SSM4, SSM5 and SSM6 methods. After creating the data set for each entity, we applied the additional pre-processing steps described by Azar [2009]: (1) application of stemming; (2) removal of *stopwords* using the list provided by *WordNet*; (3) removal of words that occur less than three times in the collection, which were considered of little importance.

To evaluate the quality of the classifiers in this work we use the traditional accuracy metric [Manning and Schütze 1999], which is the percentage of test documents that were correctly classified. The experiments done with each classifier used a 5-fold cross validation [Mitchell 1997] of the dataset, and the results presented in the next section refer to the average accuracy of 5-folds.

5.2 Results

In this section, we present the results of the previously described experiment. We first present our baseline choosing criteria, then we show the results of applying our methods.

5.2.1 *Baseline.* We adopted as baseline the supervised method proposed by Azar [2009], which also explores the polarity analysis in the financial domain. We evaluated this method using our dataset of documents annotated with their global polarities as training set. This is a configuration more similar to that used by Azar as, in practice, he considers that the documents have one single target entity and, consequently, one single polarity. Moreover, based on the idea presented in [Ward et al. 2011], we separated, for each entity E, the documents that cite E, leading to five document sets. Using these five sets of documents, we built five models, one for each entity. Therefore, the model associated with the entity E uses the pages that cite E as training examples. This corresponds to the method DbM, described in Section 4.1.

We also did a comparison with the lexical, non-supervised method, proposed by Moilanen and Pulman [2009], which we call in this paper M&P09. Although the authors did not make their sentiment grammar available, which is necessary to implement this method, we have acquired the license for the TheySay API[12] which enables polarity analysis, and we did tests with the dataset used in our work. The method was also evaluated according to the polarities of the entities, as DdM. Table III shows the results obtained for this experiment. Observe that we evaluated the methods according to the polarities of the entities.

Compared to the global method, the M&P09 had a worse performance than expected. That is due to the fact that the method is biased towards negative and positive polarities. As most of the documents

[12]http://www.theysay.io/

Table III. Method proposed by Azar trained with global polarities (Global), entity polarities (DbM strategy) and lexical method M&P09, evaluated by entity.

Entities	Global (%)	M&P09 (%)	Gain (%)	DbM (%)	Gain (%)
Apple	55.87	47.52	-14.95	58.49	4.68
Google	54.17	57.48	6.11	61.42	13.38
Samsung	52.46	48.21	-8.10	61.11	16.48
Microsoft	54.42	54.28	-0.26	71.54	31.45
Nokia	45.72	45.90	0.39	57.60	25.98

from the dataset are neutral, its performance was poor. In private contact with the authors, we have been informed that, on its default settings, the M&P09 focus on more general entities from the real world, which are usually are positive (for example, win in a soccer game or have a holiday) or negative (such as becoming sick or hospitalized), always trying to find even the weakest signs that indicate such polarities. Based on this first result, we can conclude that purely lexical methods are not really robust to determine the polarity of individual entities.

Regarding the method proposed by Azar we observe that, as expected, using multiple models, one by entity, is better than using one single global model. The smallest gain obtained was for Apple. In general, as the number of documents shrinks, the gain grows. This suggests that learning a model for each entity led to a better understanding of the patterns of less popular entities, escaping from the characteristic vices of the most popular entities. For example, from the collection of one thousand documents, Apple is the most cited. However, many citations are due to the fact that it is the manufacturer of iPhone or iPad. From the results we can conclude that a method based on multiple models, which considers the individual entity polarities, is better than a method that infers the polarity globally. Based on the results, in the next section, we use DbM as our baseline.

5.2.2 *Methods Based on Segmentation by Sentence.* In this section, we are interested in evaluating our methods that use just the sentence sets that cite the entities we are interested in as training examples. This was done in order to verify what is the accuracy of the classifiers without considering the noise added by pages that disregard the considered entity. In particular, we are interested in comparing all of our sentence segmentation methods proposed in Section 4.1.1 with DbM, where the documents are not segmented, in order to verify if this segmentation approach can lead to gains in accuracy.

The results of those experiments are presented in Table IV. In this table, the lines represent the DbM baseline and the methods based on sentences using string matching (SSM1 to SSM3) and Anaphora Resolution (SSM4 to SSM6). The results represent the accuracy obtained by the models on the task of classifying the documents as positive, negative and neutral to the entities Apple, Google, Samsung, Microsoft and Nokia. To each one of the methods, we also presented the gain (or loss) in relation to the baseline.

We observed in Table IV that SSM2 (sentences that do not cite any entity are discarded) presented greater gains in relation to DbM with regards to all entities, except by Microsoft. A similar behavior

Table IV. Method based on segmentation by sentence (SSM1, SSM2, SSM3, SSM4, SSM5 e SSM6) versus method based on document (DbM).

Method	Apple%	G%	Google%	G%	Samsung%	G%	Microsoft%	G%	Nokia%	G%
DbM	58.49	-	61.42	-	61.11	-	71.54	-	57.60	-
SSM1	57.13	-2.33	61.67	0.41	59.63	-2.42	69.76	-2.49	53.60	-6.94
SSM2	61.95	**5.92**	62.86	**2.34**	63.79	**4.39**	69.07	-3.45	64.00	**11.11**
SSM3	57.66	-1.43	62.62	1.95	59.95	-1.90	68.35	-4.46	56.80	-1.39
SSM4	53.47	-8.59	59.29	-3.47	61.65	0.88	67.64	-5.45	55.20	-4.17
SSM5	59.66	**2.01**	64.05	**4.28**	64.01	**4.74**	69.06	-3.47	60.00	**4.17**
SSM6	54.73	-6.43	60.95	-0.76	60.16	-1.56	67.64	-5.45	51.20	-11.11

Table V. Method based on segmentation by sentence (SSM2) versus methods based on detection of subjectivity (M&P09-H e SSM2-H)

Entity	SSM2 (%)	M&P09-H (%)	Gain (%)	SSM2-H (%)	Gain (%)
Apple	61.95	59.12	-4.57	61.11	-1.36
Google	62.86	62.14	-1.14	63.33	0.75
Samsung	63.79	59.05	-7.43	64.09	0.48
Microsoft	69.07	66.55	-3.65	66.90	-3.14
Nokia	64.00	61.60	-3.75	61.60	-3.75

was observed in the corresponding method based in Anaphora Resolution (SSM5). Despite anaphora resolution being a sophisticate NLP technique, the results for the strategies that use Anaphora Resolution, overall, were not the better in comparison with their non-anaphora counterparts.

In general, from these experiment sets, we can conclude that methods that use segmentation by sentence are not better than methods based in documents, even when using complex NLP techniques as Anaphora Resolution. The only segmentation methods that were consistently better, the one that discard sentences that do not cite entities, presented small gains. For all the other cases, the segmentation methods by documents presented satisfactory results for the task of polarity classification with multiple entities. As among the discarding strategies, SSM2 and SSM5, SSM2 has smaller processing cost, it will be used as basis for comparison in the next section.

5.2.3 *Methods Based on Subjectivity.* In this section, we applied a two steps hierarchical classification technique. On the first step, the classifier separates the neutral documents from the non-neutral. In other words, the first step corresponds to the task of subjectivity detection. In the second step, the documents classified as non-neutral are separated into positive or negative. Thus, the second step corresponds to a binary polarity classification.

For the first step, we used the same method applied in SSM2, trained for neutral and non-neutral classes. For the second step, we used two different classifiers: (a) the same method applied in the first step, trained with positive and negative classes, and (b) the non-supervised method M&P09. As seen in Section 5.2.1, M&P09 was not competitive since it was biased towards non-neutral sentiments. As will be seem in the experiments, it actually turns out to be a viable option as classifier for the second step.

Table V presents the results of the methods presented in this section with its respective gains in relation to the method SSM2. The hierarchical method, supervised in both steps, is called SSM2-H. The hierarchical method with a non-supervised second step is called M&P09-H.

As can be seen in Table V, in general, the hierarchical strategies were not capable of producing better results in comparison to the non-hierarchical ones. In the case of M&P09-H, however, the observed error significantly shrinks in relation to the original results, which was the expected behavior. Nevertheless its results were inferior to the ones produced by supervised strategies. Note however that any non-supervised effort has the great advantage of not requiring manually labeled examples.

Regarding SSM2-H, the results obtained were underwhelming. When analyzing these results in detail, we realized that SSM2-H achieves better performance in the classification step, with an improved recognition of neutral cases. Nonetheless, its performance in the second step is visibly worse, since even though the classifier of the first step is better in detecting neutral cases, it still makes many errors. As a result, many neutral sentences are not filtered out and go to the second step. This behavior is specially visible in the experimental dataset due to its polarity distribution, which is quite biased towards the neutral polarity, in almost 70% of the documents.

These results led us to think that a semi-hierarchical method is more appropriate for this problem. Instead of using a binary polarity classifier in the second step, it would be better to use again a

Table VI. Method based on segmentation by sentence (SSM2) versus a semi hierarchical method based on detection of subjectivity (SSM2-SH). Results obtained with the method DbM are also shown, for reference.

Entidade	DbM (%)	SSM2 (%)	SSM2-SH (%)	Gain (%) over DbM	Gain (%) over SSM2
Apple	58.49	61.95	62.89	7.52	1.52
Google	61.42	62.86	64.76	5.44	3.03
Samsung	61.11	63.79	65.58	7.31	2.80
Microsoft	71.54	69.07	70.11	-2.00	1.50
Nokia	57.60	64.00	65.60	13.89	2.50

classifier for the three classes. This is a way to reduce the error of the second step, once some neutral cases can still be observed. We evaluate this new classifier, called SSM2-SH, in Table VI.

In this approach, the gains over the segmentation method, DbM, are more visible, specially if we consider the most common entities. In relation to SSM2 based methods, SSM2-SH obtained better gain levels in comparison to SSM2-H, but those were still modest. Although in the second step we built a model that is trained with examples from the three classes, this model was obtained using the whole collection. This is not the best strategy since we verified that, in the second step,the number of neutral cases corresponds to only two thirds of the number present the whole dataset. In other words, the distribution of neutral cases in the whole collection does not reflect the distribution in the second step, which might lead to distortions in the trained models.

A way to fix this problem would be to use a training collection to estimate what is the final distribution of positive, negative and neutral documents after the application of the classifier in the first step. This could be done with a cross-validation of the training cases in order to determine in which instances the classifier of the first step would be considered non-neutral. These instances could be used as training for the new second step classifier.

6. CONCLUSIONS AND FUTURE WORK

In this article, we studied how to improve the performance of classification models in the task of polarity classification in financial documents with multiple entities. Our results demonstrated that an approach based on multiple models is able to obtain significant gains over an approach based on global model in the polarity classification task with multiple entities. The segmentation of the document into sentences that mention the entities and the adoption of a hierarchical strategy also achieved gains, although those were modest.

Although the results of the methods using the hierarchical approach had not lead to satisfactory gains, this is an interesting approach for the polarity classification task, specially if applied to scenarios where the dataset provides a more balanced distribution of polarities. We also noted that even a sophisticated unsupervised lexical method is not as effective as a completely supervised method in the polarity classification task in documents with multiple entities. However, they become more competitive when combined with a supervised method in an hierarchical approach. From a practical point of view, they will be always an alternative to be considered since they do not depend on labeled data, which is extremely hard to obtain in many scenarios.

In this work, we assumed that the polarities of the different entities are independent, what probably might not occur very often in real life. For example, in our collection, when a document is positive to Apple, it is more likely to be negative to Samsung than a random document. Thus, in the future, we intend to investigate techniques that consider the mutual influence of the entity polarities. In particular, we will study techniques discussed in [Dembczyński et al. 2010], such as stacking and multi-variate regression. Another future work is the more detailed investigation of new methods for semi-hierarchical classification, where the definition of the training samples for the more specific steps consider the error rate of the more general steps. In addition, we intend to verify our conclusions in

collections from other domains and to study hybrid methods based on multiple models to deal with a large number of entities.

REFERENCES

Abdul-Mageed, M., Diab, M. T., and Korayem, M. Subjectivity and Sentiment Analysis of Modern Standard Arabic. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: human language technologies*. Portland, USA, pp. 587–591, 2011.

Abdul-Mageed, M., Kübler, S., and Diab, M. SAMAR: a system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Republic of Korea, pp. 19–28, 2012.

Araújo, M., Gonçalves, P., and Benevenuto, F. Measuring Sentiments in Online Social Networks. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. Salvador, Brazil, pp. 97–104, 2013.

Azar, P. D. *Sentiment Analysis in Financial News*. Ph.D. thesis, Harvard University, 2009.

Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval: the concepts and technology behind search*. Addison-Wesley, 2011.

Becker, K. and Tumitan, D. Introdução à Mineração de Opiniões: conceitos, aplicações e desafios. In *Proceedings of the Brazilian Symposium on Databases, Lectures*. Recife, Brazil, pp. 27–52, 2013.

Bollen, J., Mao, H., and Zeng, X. Twitter Mood Predicts the Stock Market. *Journal of Computational Science* 2 (1): 1–8, 2011.

Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (3): 27:1–27:27, 2011.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. On Label Dependence in Multi-label Classification. In *Proceedings of the International Workshop on Learning from Multi-label Data*. Haifa, Israel, pp. 5–12, 2010.

Ferreira, J. Z., Rodrigues, J., Cristo, M., and de Oliveira, D. F. Multi-entity Polarity Analysis in Financial Documents. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. João Pessoa, Brazil, pp. 115–122, 2014.

Gryc, W. and Moilanen, K. Leveraging Textual Sentiment Analysis with Social Network Modelling. In *Proceedings of the "From Text to Political Positions" Workshop*. Amsterdam, Netherlands, pp. 47–70, 2010.

Im, T. L., San, P. W., On, C. K., Alfred, R., and Anthony, P. Analysing Market Sentiment in Financial News Using Lexical Approach. In *Proceedings of the IEEE Conference on Open Systems*. Sarawak, Malaysia, pp. 145–149, 2013.

Liu, B. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, System Demonstrations Track*. Baltimore, USA, pp. 55–60, 2014.

Mitchell, T. M. *Machine Learning*. McGraw-Hill, 1997.

Moilanen, K. and Pulman, S. Multi-entity Sentiment Scoring. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria, pp. 258–263, 2009.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña-López, L. A. Ranked WordNet Graph for Sentiment Polarity Classification in Twitter. *Computer Speech & Language* 28 (1): 93–107, 2014.

Mourad, A. and Darwish, K. Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Atlanta, USA, pp. 55–64, 2013.

Ng, V. Supervised Noun Phrase Coreference Research: the first fifteen years. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 1396–1411, 2010.

Pang, B. and Lee, L. A Sentimental Education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Barcelona, Spain, pp. 1–8, 2004.

Poesio, M., Ponzetto, S., and Versley, Y. Computational Models of Anaphora Resolution: a survey. http://cswww.essex.ac.uk/poesio/papers.html, 2011.

Refaee, E. and Rieser, V. Subjectivity and Sentiment Analysis of Arabic Twitter Feeds with Limited Resources. In *Proceedings of the Workshop on Free/Open-source Arabic Corpora and Corpora Processing Tools*. Reykjavík, Iceland, pp. 16–21, 2014.

Romanyshyn, M. Rule-based Sentiment Analysis of Ukranian Reviews. *International Journal of Artificial Intelligence & Applications* 4 (4): 103–111, 2013.

Schumaker, R. P. and Chen, H. Textual Analysis of Stock Market Prediction Using Breaking Financial News: the AZFin text system. *ACM Transactions on Information Systems* 27 (2): 12:1–12:19, 2009.

Ward, C. B., Choi, Y., Skiena, S., and Xavier, E. C. Empath: a framework for evaluating entity-level sentiment analysis. In *International Conference & Expo on Emerging Technologies for a Smarter World*. Hauppauge, USA, pp. 1–6, 2011.

Witten, I. H., Frank, E., and Hall, M. A. *Data Mining: practical machine learning tools and techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2011.