Undefined 1 (2014) 1–5
IOS Press

An Abstract Argumentation Approach for the Prediction of Analysts' Recommendations following Earnings Conference Calls

Andrea Pazienza [a,*], Davide Grossi [b], Floriana Grasso [c], Rudi Palmieri [d], Michele Zito [c] and Stefano Ferilli [a]

[a] *Dipartimento di Informatica, University of Bari Aldo Moro, Via E. Orabona, 4, Bari, Italy*
E-mail: {andrea.pazienza,stefano.ferilli}@uniba.it
[b] *Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands*
E-mail: d.grossi@rug.nl
[c] *Computer Science Department, University of Liverpool, Ashton St., Liverpool L69 3BX, UK*
E-mail: {floriana,michele}@liverpool.ac.uk
[d] *Communication and Media Department, University of Liverpool, 19 Abercromby Square, Liverpool, L69 3BX, UK*
E-Mail: rudi.palmieri@liverpool.ac.uk

Abstract. Financial analysts constitute an important element of financial decision-making in stock exchanges throughout the world. By leveraging on argumentative reasoning, we develop a method to predict financial analysts' recommendations in earnings conference calls (ECCs), an important type of financial communication. We elaborate an analysis to select those reliable arguments in the Questions & Answers (Q&A) part of ECCs that analysts evaluate to estimate their recommendation. The observation date of stock recommendation update may variate during the next quarter: it can be either the day after the ECC or it can take weeks. Our objective is to anticipate analysts' recommendations by predicting their judgment with the help of abstract argumentation. In this paper, we devise our approach to the analysis of ECCs, by designing a general processing framework which combines natural language processing along with abstract argumentation evaluation techniques to produce a final scoring function, representing the analysts' prediction about the company's trend. Then, we evaluate the performance of our approach by specifying a strategy to predict analysts recommendations starting from the evaluation of the argumentation graph properly instantiated from an ECC transcript. We also provide the experimental setting in which we perform the predictions of recommendations as a machine learning classification task. The method is shown to outperform approaches based only on sentiment analysis.

Keywords: Argumentation, Natural Language Processing, Sentiment Analysis, Machine Learning

1. Introduction

Earnings conference calls are one of the most important types of financial communication. As soon as their periodic results are announced (typically quarterly earnings reports), publicly-listed corporations organise a teleconference, or webcast, in which the financial results are presented to and discussed with financial analysts. The main participants to this regular communicative event are the corporate executive managers (the Chief Execu-

*Corresponding author. E-mail: andrea.pazienza@uniba.it

tive Officer and the Chief Financial Officer in particular) and financial analysts, whose institutional role is that of scrutinising corporate statements and formulate recommendations for investors who own or may wish to buy the shares of the company. ECCs follow the release of the company's quarterly earnings announcements and are divided in two main parts [1]: first, corporate executives present the period results with analysts put in a listen-only mode (presentation part); subsequently, analysts take the line and ask questions to which corporate representatives reply immediately. Often, a question turn includes several questions which are dealt with by different corporate executives. Follow-up questions are possible. An independent operator manages the call.

As [2,3] explain, the participation in an ECC is motivated by both informative and rhetorical objectives. Analysts are interested in getting valuable information that can help them construct reliable recommendations, which in turn help investors in making more accurate investment decisions (buy, hold or sell shares). At the same time, companies have an interest in releasing information and clarifying matters because a better informed market leads to a lower cost of capital for them. ECCs are in fact forms of voluntary, not compulsory, disclosure which by definition are motivated by strategic objectives rather than compliance duties. Obviously, corporate managers strive to persuade analysts to positively evaluate the firm results and, by linking results to managerial actions, to induce a positive impression about their image and reputation. This makes ECC an inherently rhetorical genre where a variety of communicative strategies can support managerial objectives.

The linguistic content of ECC has been studied in financial accounting studies and, in more recent years, by scholars in communication disciplines, such as linguistics, argumentation and rhetoric (for a systematic literature review, see [4]). The former have been particularly interested in determining the informative value of these disclosure events, with some evidence of the Questions & Answers (Q&A) part being incrementally informative over the presentation part and the presentation part being incrementally informative over the earnings announcement preceding the call [5]. However, less evidence exists on the actual causes and sources of such informativeness. Taking a discourse-analytics perspective, [2] hypothesises that the presence of argumentation acts as a relevant factor making the content of ECC informationally useful and price sensitive. While their analysis is limited to the Q&A part without examining the possible impacts on market events (e.g. stock prices, volatility, volumes, analyst recommendations), numerous argumentative patterns are brought to light which suggest argumentation plays a decisive role in this context.

We are interested in the argumentative and dialogical patterns arising in ECCs, and the present paper could be broadly placed within the recent body of research in Argument Mining (see [6] for an excellent introduction), and in particular related to works such as [7,8,9,10], or to opinionated claim mining [11]. This paper addresses however the less explored issue of the evaluation of arguments, in terms of their persuasive effect, recognised as a challenge by many [12]. In this sense, this paper is in the spirit of works such as [13,14,15], though it provides a more operational and pragmatic evaluation measure, derived from the context we explore.

Related to our approach is also work aiming at providing high-level representations of debate interaction such as [16] or [17], which developed and studied graph-theoretic representations of parliamentary debates in the Netherlands and, respectively, the UK. Inasmuch as our work bridges computational abstract models of argument, and argumentation in real world domains, it contributes to a wider on-going research effort aiming at making argumentation technology significant for applications (cf. [18]).

In this paper we propose a novel approach to the analysis of ECC, especially during their Q&A component, which is grounded in computational argumentation. We focus on the type of interaction between analysts and corporate representatives (essentially, who talks to whom and with what tone, cf. [16,17]) and are interested in studying whether, and if so how, this interaction has an effect on analysts' recommendations. We collected ECC transcripts concerning 10 major companies in the 2007-12 period. In line with [19], to model the argumentative interaction occurring in the Q&A of these ECC we used bipolar weighted argumentation frameworks (BWAF, [20]) where we considered as basic units of analysis—or, 'arguments'—each intervention by an analyst or corporate representative, and provided specific NLP-based met-

rics to recognise relations of attack or support among these interventions. Once an ECC has been modeled as a BWAF we single out 'strong' arguments in the ECC using a novel ranking-based semantics specifically developed for the analysis of ECCs. Our hypothesis is that such 'strong' arguments carry more weight in influencing the analysts' perception of the ECC. The obtained BWAFs, their analysis, together with data on analysts' recommendations, as well as financial performance indicators for the relevant companies have then been used to create a novel dataset amalgamating argumentative and financial information. To the best of our knowledge, this is the first data set of its kind in the computational argumentation literature, covering financial as well as argumentative features. With this data set, using off-the-shelf machine learning techniques, we show that incorporating argumentative features in the learning task improves prediction of analysts' recommendation over techniques using only sentiment analysis (e.g., [21]). This finding corroborates the hypotheses put forth in [2] that argumentative structure carries informational value for analysts in ECCs, and in [22] that an abstract model of the local sentiment flow captures the overall argumentation regarding global sentiment.

The paper is organised as follows. Next section recalls the background and basic concepts on abstract argumentation theory useful for our analysis, Section 3 develops the method as a general processing framework divided into four phases: natural language processing model, bipolar weighted graph instantiation, semantics evaluation and tone-based evaluation. Section 4 performs the experimental setting to validate our approach. Finally, Section 5 concludes the paper.

2. Background

The section introduces the toolbox from abstract argumentation theory that will be later used in our analysis of ECCs.

2.1. Bipolar Weighted Argumentation Frameworks

Dung's Argumentation Frameworks [23] (in short, AF) play a special role in the representation of argument interaction: arguments are nodes in a directed graph, edges in such graphs represent attack relations among arguments, and graph-theoretic notions (e.g., stable sets or kernels) acquire natural argumentative interpretations as 'reasonable'—with respect to different intuitive standards—sets of arguments. An argumentation semantics is the formal definition of a method ruling the argument evaluation process. The most basic concepts shared by all argumentation semantics in the literature are *conflict-freeness* (i.e., an attacking and an attacked argument can not stay together) and *defense* (i.e., replying to every attack with a counterattack). In this way, an attacker a of an argument b is an argument at the beginning of an odd-length path, while a defender a of b is an argument at the beginning of an even-length path. Dung's original formalism for abstract argumentation has been extended along many lines giving rise to a large and thriving literature in AI (see [24,25] for an overview). The extensions that are relevant for the purpose of this paper are two: bipolar argumentation frameworks, and weighted argumentation frameworks.

A Bipolar AF (BAF) [26] is an extension of Dung's AF in which two kinds of interactions between arguments are possible: the attack relation and the support relation. A BAF can be represented by a directed graph in which two kinds of edges are used, in order to differentiate between the two relations. In BAFs, new kinds of attack emerge from the interaction between the direct attacks and the supports: there is a *supported attack* for an argument b by an argument a iff there is a sequence of supports followed by one attack, while, there is an *indirect attack* for an argument b by an argument a iff there is an attack followed by a sequence of supports. In particular, we assume to say that a supports b if there is a sequence of direct supports from a to b. Taking into account sequences (i.e., paths) of supports and attacks it is possible to revise Dung's definitions of acceptability applying to sets of arguments.

A Weighted AF (WAF) [27] is another extension of Dung's AF in which attacks between arguments are associated with a weight, indicating the relative strength of the attack. Note that allowing 0-weight attacks is counter-intuitive since it can be interpreted as absence of attack relation. In this framework, some inconsistencies are tolerated in subsets S of arguments, provided that the sum of the weights of attacks between argu-

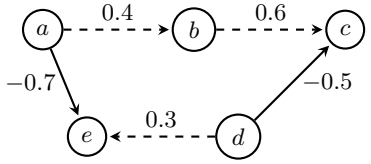

Fig. 1. G_1: Example to illustrate BWAF

ments of S does not exceed a given inconsistency budget $\beta \in \mathbb{R}_*^+$. The meaning is that attacks up to a total weight of β are neglected. Dung's argument systems assume an inconsistency budget of 0, while, by relaxing this constraint, WAFs can achieve more solutions.

A Bipolar Weighted AF (*BWAF*) [28] incorporates both above generalizations of Dung-style AFs. The idea behind it is to allow not only weighted attack relations between abstract arguments, but also weighted support relations. This is achieved by assigning to each relation a weight which can be positive or negative.

Definition 1. *A BWAF is a triplet* $G = \langle \mathcal{A}, \hat{\mathcal{R}}, w_{\hat{\mathcal{R}}} \rangle$, *where* $\mathcal{A}$ *is a finite set of arguments,* $\hat{\mathcal{R}} \subseteq \mathcal{A} \times \mathcal{A}$ *and* $w_{\hat{\mathcal{R}}} \colon \hat{\mathcal{R}} \mapsto [-1, 0[\cup]0, 1]$. *Attack relations are defined as* $\hat{\mathcal{R}}_{att} = \{\langle a, b\rangle \in \hat{\mathcal{R}} \mid w_{\hat{\mathcal{R}}}(\langle a, b\rangle) \in [-1, 0[\ \}$ *and support relations as* $\hat{\mathcal{R}}_{sup} = \{\langle a, b\rangle \in \hat{\mathcal{R}} \mid w_{\hat{\mathcal{R}}}(\langle a, b\rangle) \in]0, 1]\ \}$.

Given two arguments $a, b \in \mathcal{A}$ *and a path* $\langle a, x_1, x_2, \ldots, x_n, b\rangle$ *from a towards b, then:*

- *a* bw-defends *b if the product of weights* $w_{\hat{\mathcal{R}}}(\langle a, x_1\rangle) \cdot w_{\hat{\mathcal{R}}}(\langle x_1, x_2\rangle) \cdot \ldots \cdot w_{\hat{\mathcal{R}}}(\langle x_n, b\rangle)$ *is positive.*
- *a* bw-attacks *b if the product of weights* $w_{\hat{\mathcal{R}}}(\langle a, x_1\rangle) \cdot w_{\hat{\mathcal{R}}}(\langle x_1, x_2\rangle) \cdot \ldots \cdot w_{\hat{\mathcal{R}}}(\langle x_n, b\rangle)$ *is negative.*

As you can see in Figure 1, a BWAF can be represented as a directed graph whose nodes represent arguments, relations represent attacks (with normal arcs) and supports (with dashed arcs), and weights represent the relative strength of relations. In what follows we will often abuse our notations and use G to denote the whole BWAF or its underlying directed graph. BWAFs introduce a generalised notion of defense based on the concept of transitivity of a *multiplication rule* in which: (i) it is loose the basic Dung's notion in which even-length paths of attacks means a defense (i.e., *the attack of an attack is a defense*); (ii) BAF's notions of indirect attack and supported attack are both covered by a single definition.

Example 1. *In the BWAF* $G_1 = \langle \mathcal{A}_1, \hat{\mathcal{R}}_1, w_{\hat{\mathcal{R}}_1} \rangle$ *shown in Figure 1, we have:*
$\mathcal{A}_1 = \{a, b, c, d, e\}$, $\hat{\mathcal{R}}_1 = \{\langle a, b\rangle, \langle b, c\rangle, \langle a, e\rangle, \langle d, e\rangle, \langle d, c\rangle\}$
where $w_{\hat{\mathcal{R}}_1}(\langle a, b\rangle) = 0.4$, $w_{\hat{\mathcal{R}}_1}(\langle b, c\rangle) = 0.6$, $w_{\hat{\mathcal{R}}_1}(\langle a, e\rangle) = -0.7$, $w_{\hat{\mathcal{R}}_1}(\langle d, e\rangle) = 0.3$, $w_{\hat{\mathcal{R}}_1}(\langle d, c\rangle) = -0.5$,
such that $\hat{\mathcal{R}}_{att} = \{\langle a, e\rangle, \langle d, c\rangle\}$, $\hat{\mathcal{R}}_{sup} = \{\langle a, b\rangle, \langle b, c\rangle, \langle d, e\rangle\}$.

2.2. Ranking-based Semantics for BWAFs

BWAFs will be used in this paper as an abstract representation of argumentative interaction in the Q&A of an ECC. So once an ECC is represented as a BWAF, we need a computationally feasible method to automatically analyze the BWAF in order to single out 'influential' interactions—or 'strong' arguments—in the framework. This calls naturally for the application, to BWAFs, of ranking-based semantics [29] methods. Intuitively, a ranking-based semantics determines, for any framework—in our case BWAFs—a ranking of the available arguments in the form of a pre-order (reflexive and transitive relation). In our case, given that BWAFs will be extracted from real data, we want the ranking process to be computationally viable. This rules out the application of existing ranking-based semantics for BWAFs, the so-called *sp*-semantics [20]. In fact, we may exploit this semantics due to its ability to deal with weighted cycles by exploring all the possible paths (with eventually cycles and sub-cycles) between any pair of nodes in the graph, but for large graphs this may result computationally expensive [30].

Instead, for the purpose of this paper, we leverage matrix algebra methods, recently addressed in [31], to exploit a particular approach to argument ranking in BWAFs, which we refer to as *Laplacian Ranking semantics*. We do not claim this semantics to be of general applicability for the analysis of argumentation, but rather to be an effective tool for the analysis of the specific form of argumentation which is the focus of this paper.

2.2.1. Laplacian Semantics

Spectral graph theory provides techniques that apply the theory of linear maps (in particular, eigenvalues and eigenvectors) to matrices that

do not represent geometric transformations, but rather some kind of relationship between entities. It studies the properties of graphs via the eigenvalues and eigenvectors of their associated graph matrices: the adjacency matrix and the graph Laplacian and its variants. In the following we consider the possible benefits of adopting spectral linear algebra methods as a tool for analyzing argumentation structures.

Mathematically speaking, studies in Abstract Argumentation semantics are concerned with the properties of numerical measures on directed graphs. Matrix theory is an important field of Linear Algebra used in particular for representing and handling graphs. Given a directed graph G on n nodes, the adjacency matrix of G is an $n \times n$ matrix $\boldsymbol{A}_G$ whose entries $(A_G)_{ij}$ (for $1 \leq i, j \leq n$) equal 1 (resp. 0) whenever a directed edge from i to j is present (resp. not present) in the graph G. We will use the simpler notations $\boldsymbol{A}$ and A_{ij}, when the graph G is clear from the context, and no ambiguity arises. BWAFs lend themselves naturally to a generalization of this type of matrix representation.

Definition 2. *Let $G = \langle \mathcal{A}, \hat{\mathcal{R}}, w_{\hat{\mathcal{R}}} \rangle$, where $\mathcal{A}$ be a BWAF with weights in the interval $[-1, 0[\cup]0, 1]$, and $|\mathcal{A}| = n$. Then, the* Signed Weighted Argumentation Matrix *(in short,* Argumentation Matrix*) of G is a $n \times n$ matrix $\boldsymbol{M}_G$ such that for any two arguments $a_i, a_j \in \mathcal{A}$:*

$$(M_G)_{ij} = \begin{cases} w_{\hat{\mathcal{R}}}(\langle a_i, a_j \rangle) & \text{if } \langle a_i, a_j \rangle \in \hat{\mathcal{R}} \\ 0 & \text{otherwise} \end{cases}$$

For simple directed graphs, the powers of the adjacency matrix can be used to count the number of walks (i.e. directed paths) in the given graph. More specifically, if $\boldsymbol{A}^k$ is the k^{th} power of $\boldsymbol{A}$, then $(A^k)_{ij}$ gives the number of walks from i to j of length k. In BWAFs, matrix multiplication can be used in the same way. If the weight of a walk is defined as the product of the weights of the arcs in the walk, then the sum of the weights of all walks from i to j with length k will be given by $(M^k)_{ij}$. Regarding the complexity, given that these kind of matrix are diagonalizable, if $\boldsymbol{A}^k = \boldsymbol{P}^{-1}\boldsymbol{D}^k\boldsymbol{P}$, with diagonal $\boldsymbol{D}$, then the k-th power of $\boldsymbol{A}$ can be computed by just taking each element of the diagonal (each eigenvalue of $\boldsymbol{A}$) to the k-th power [32].

Critically, an alternative matrix representation of a BWAF makes it possible to obtain explicit numerical information about the effect of an argument i over an argument j, through defense or attack paths. Such representation is called the *Justification Matrix* (of the underlying BWAF G). Let $(\boldsymbol{J}_{Gn})_{n\in\mathbb{N}}$ be a sequence of matrices in which the nth term is defined as:

$$\boldsymbol{J}_{Gn} = \sum_{k=1}^{n} {\boldsymbol{M}_G}^k \tag{1}$$

(as usual we omit the subscript G whenever it is possible without compromising the clarity of the presentation). Entry $(J_n)_{ij}$ is the accumulated sum of the weights of all paths of length up to n between argument i and argument j where paths of (indirect) attacks contribute negatively and paths of (indirect) defenses contribute positively. Hence, the interpretation of a positive acceptability assessment of argument j with respect to argument i is that i supports j or "if i is accepted then so should j". On the other hand, a negative acceptability assessment of argument j with respect to argument i indicates some contradiction between the arguments and "if i is accepted then j should not be accepted". Furthermore, the jth column in $(\boldsymbol{J}_{Gn})$ gives an overview on how argument j is assessed by all arguments in the framework.

Notice that for an arbitrary BWAF G, $\boldsymbol{J}_{Gn}$ might not converge as n becomes large, but if that happens then the resulting value, which we denote by $\boldsymbol{J}_G$ is the Justification Matrix of G.

In general, for any $n \times n$ matrix $\boldsymbol{X}$ with real coefficients, the power series $\sum_{k=1}^{n} \boldsymbol{X}^k$ converges if its spectral radius (i.e. the largest absolute value of any of the eigenvalues of X) is strictly less than one [33]. As to BWAF, a simple case is that of BWAFs whose underlying graph is acyclic. If that is the case, then there is only a finite number of non-zero powers ${\boldsymbol{M}_G}^k$. More specifically we have that

$${\boldsymbol{M}_G}^{\mathrm{diam}(G)+1} = 0, \tag{2}$$

where, for the purposes of this work, $\mathrm{diam}(G)$ is the diameter of G (i.e., the length of the *longest shortest path* between any two nodes in the graph, ignoring those nodes that are not connected by

any finite directed path), and this, in turns implies that

$$\boldsymbol{J}_G = \sum_{k=1}^{\mathrm{diam}(G)} \boldsymbol{M}_G{}^k. \tag{3}$$

In BWAFs that contain cycles[1] the power series computation of $\boldsymbol{J}_G$ might not terminate. One has therefore to determine a cut-off point for the computation of the Justification Matrix. It turns out that defining $\boldsymbol{J}_G$ as in (3) also suffices for our purposes in this case. Notice, in particular, that, for any pair of arguments a_i and a_j, $(\boldsymbol{J}_G)_{ij}$ contains a value that depends on all paths from a_i to a_j. Our formalism does not distinguish between an interaction between two arguments that results in $(\boldsymbol{J}_G)_{ij} = 0$ (this may happen if there is more than one path connecting the two, and the cumulative weight of interactions between the arguments on different paths have opposite signs) and the total absence of any interaction. However, in our application, the underlying graphs of our BWAF's are strongly connected[2] and therefore any two arguments a_i and a_j are connected by a direct path. Finally, it should be noted that self-loops would pose a problem for the computation in (1). Self-loops, however, do not occur in the class of BWAFs representing ECCs.

With the definition of the Justification matrix in place, we can now proceed to the main definition:

Definition 3. *Let $G = \langle \mathcal{A}, \hat{\mathcal{R}}, w_{\hat{\mathcal{R}}} \rangle$ be a BWAF, with $|\mathcal{A}| = n$, and let $\boldsymbol{J}_G$ be the Justification Matrix of G. The* degree matrix *of G is the matrix $\boldsymbol{D}_G = diag(\deg(a_1), \ldots, \deg(a_n))$, where $a_1, \ldots, a_n \in \mathcal{A}$, and $\forall j = 1, \ldots, n$:*

$$\deg(a_j) = \sum_{i=1}^{n} (J_G)_{ij}$$

Intuitively, the degree matrix of a BWAF is a diagonal matrix which contains information about the sum of weights of the edges connected to a node.[3] In yet other words, the degree matrix $\boldsymbol{D}_G$ collects in the main diagonal the column-wise sum of its entries. Hence, we argue, it captures natural information to compare the relative 'strength' of arguments in a BWAF, since its Justification Matrix collects all the attacks and defenses for each node in the graph.

Example 2. *Consider the BWAF G_1 depicted in Figure 1. Below, $\boldsymbol{M}_{G_1}$ is its Argumentation Matrix. We can compute its Justification Matrix $\boldsymbol{J}_{G_1}$ with the power series summation of $\boldsymbol{M}_{G_1}$. Below, $\boldsymbol{M}_{G_1}{}^2$ is the 2nd power of $\boldsymbol{M}_{G_1}$. Since there is no path of length three in G_1, $\boldsymbol{M}_{G_1}{}^3$ is the zero matrix. Then, $\boldsymbol{J}_{G_1}$ is the resulting Justification Matrix of G_1. In particular, the degree matrix of G_1 is $\boldsymbol{D}_{G_1}$.*

$$\boldsymbol{M}_{G_1} = \begin{bmatrix} 0 & 0.4 & 0 & 0 & -0.7 \\ 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.5 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{M}_{G_1}{}^2 = \begin{bmatrix} 0 & 0 & 0.24 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{J}_{G_1} = \begin{bmatrix} 0 & 0.4 & 0.24 & 0 & -0.7 \\ 0 & 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.5 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{D}_{G_1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.34 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.4 \end{bmatrix}$$

So we assign an 'acceptability' degree to each argument in a BWAF G, which equals its degree in $\boldsymbol{D}_G$. It follows that the degree of an argument always lies in the interval $[-1, 1]$, so that the ranking of 0 will now *tip the scales*, meaning that rejected arguments will have a negative ranking, while ac-

[1] The BWAFs we will be studying will mostly be of this type.

[2] In fact for each a_i and a_j both $\langle a_i, a_j \rangle$ and $\langle a_j, a_i \rangle$ belong to $\hat{\mathcal{R}}$.

[3] The name 'Laplacian semantics' derives from the fact that, in graph theory, the Laplacian matrix of a graph G is given by the difference $\boldsymbol{D}_G - \boldsymbol{J}_G$.

cepted ones will have a positive ranking. Naturally, such degrees induce a total preorder.

Definition 4. *The* Laplacian ranking semantics *associates to any BWAF* $G = \langle \mathcal{A}, \hat{\mathcal{R}}, w_{\hat{\mathcal{R}}} \rangle$ *a ranking* $\succeq_G^{\deg}$ *on* $\mathcal{A}$ *such that* $\forall a, b \in \mathcal{A}, a \succeq_G^{\deg} b$ *iff* $\deg(a) \geq \deg(b)$.

Example 3. *Consider again the BWAF* G_1 *depicted in Figure 1. Given* $\boldsymbol{D}_{G_1}$*, i.e. the degree matrix of* G_1 *in which* $\deg(a) = 0, \deg(b) = 0.4, \deg(c) = 0.34, \deg(d) = 0, \deg(e) = -0.4$*, then the Laplacian ranking semantics of* G_1 *is:*

$$b \succ_{G_1}^{\deg} c \succ_{G_1}^{\deg} a \succeq_{G_1}^{\deg} d \succ_{G_1}^{\deg} e.$$

3. ECCs as BWAFs

A key objective of our analysis of ECCs consists in being able to automatically recognise which arguments are likely to be the most relevant in a given ECC transcript. To this aim we design a general processing framework, which is divided in four fundamental phases: *natural language processing model*, *bipolar weighted graph instantiation*, *semantics evaluation* and *tone-based evaluation*. The natural language processing (NLP) model is carried out to analyze the text of the ECC transcript and the graph building procedure to perform a mining task of both recognizing arguments and identifying relations between them, jointly modeling an argumentation structure that is, in this case, the BWAF.

Given an ECC to be analysed (see for instance Figure 2), we apply a processing procedure that progressively splits the Q&A part of the ECC transcript into arguments, and analyzes the sentiment of each argument. After that, we build the relations between arguments that are exploited to generate a BWAF. Each relation is a couple ⟨question, answer⟩ or ⟨answer, question⟩ whose weight represents the degree of attack/support between them. The resulting BWAF is exploited to generate a ranking of acceptability for arguments in which they can be evaluated as either accepted or rejected with a different degree. Finally, we design a procedure to evaluate the final trend of the ECC with a scoring value. Such aggregation measure, derived from the evaluation of the BWAF, will be used to predict the recommendation rating of the analysts involved.

3.1. NLP Model

Each paragraph in the transcript is assumed to be a single (abstract) argument. The NLP Model is in charge of extracting the sentiment of each argument. To quantify sentiment, we initially need to determine which arguments are positive or negative. This is accomplished by exploiting a dictionary of words. The *Stanford CoreNLP* toolkit [34] provides a set of natural language analysis tools, including the sentiment analysis (SentimentAnnotator) and various programs which support it. Such a model can be used to analyze text as part of StanfordCoreNLP by adding "sentiment" to the list of annotators.

Stanford CoreNLP (SC for short) is therefore exploited to extract sentiment from arguments. There is a drawback however: its sentiment dictionary uses only a standard English dictionary to classify words as negative or positive. This is not fully exploitable when the text to be analysed is finance-related. For instance, if an argument exposed in an ECC transcript contains a disproportionate number of terms like "shortfall" and "decline" then it is reasonable to think that its sentiment is negative. To solve this problem we use a financial dictionary [35] (LM for short) with customised lists of negative and positive words specific to the accounting and financial domain. LM provides a clear demonstration that applying a general sentiment word list to accounting and finance topics can lead to a high rate of misclassification. For example, words like "mine", "cancer", "tire" or "capital" are often used to refer to a specific industry segment. These words are not predictive of the tone of documents or of financial news and simply add noise to the measurement of sentiment and attenuate its predictive value. [4]

For the above reasons, the overall tone of each argument is computed by averaging the tone coming from both the SC and LM dictionaries. The combination of SC and LM dictionaries is exploited and then combined to accomplish the pos-

[4] Although not relevant for our study, the LM dictionary has the additional benefit of covering dimensions of interest beyond the traditional dichotomy positive/negative. Worth mentioning are the Uncertainty word list that attempts to measure the general notion of imprecision (without an explicit reference to risks), and the Litigiousness word list that may be used to identify potential legal problem situations.

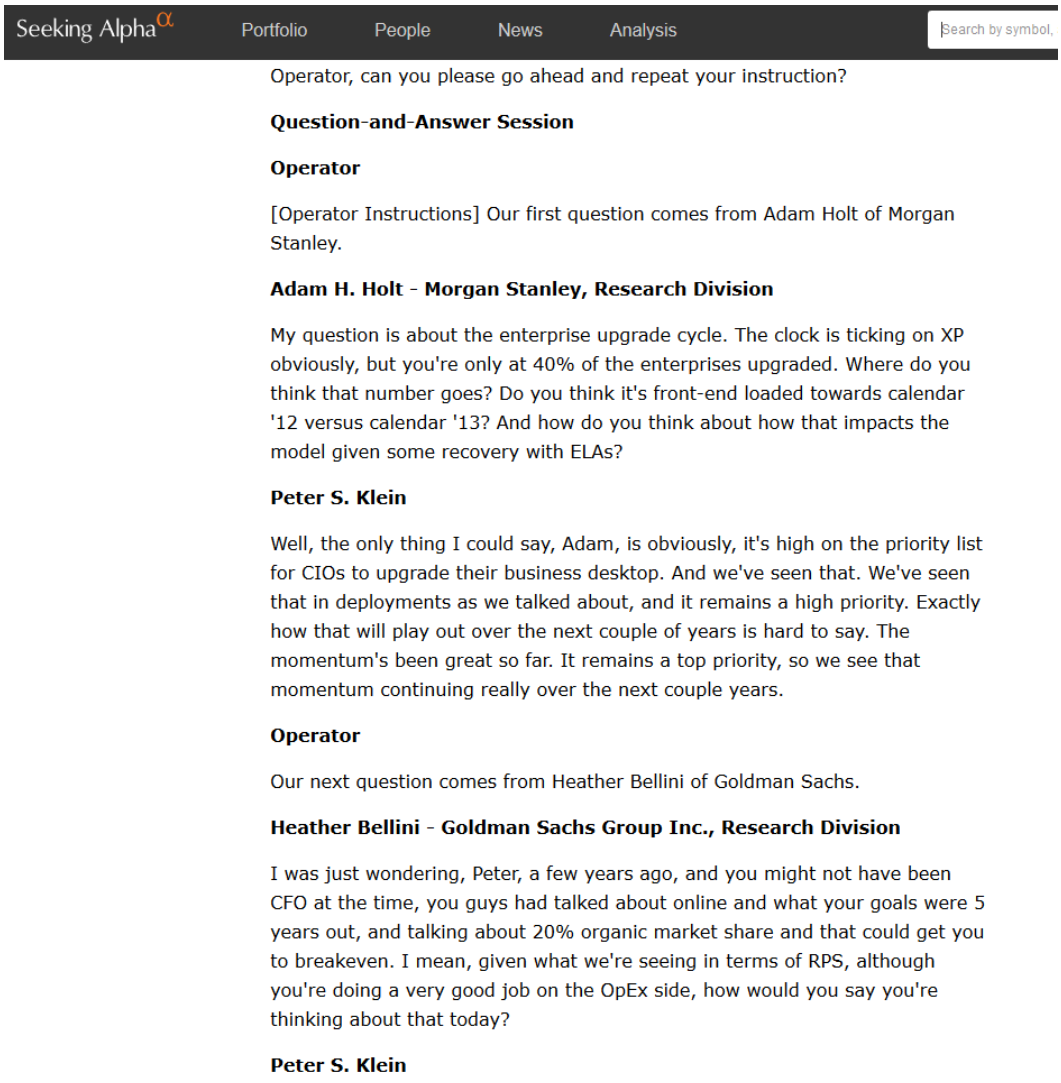

Fig. 2. Seeking Alpha web page of the ECC transcript (Q&A part only) of Microsoft Corp. in the 3rd quarter of 2012.

itive and negative word frequencies into the sentiment, or *tone*, of the arguments. The mean between SC and LM is required to capture both the general discourse made in English (by SC) and "adjusted" by LM to better remark the sentiment for words coming from financial vocabulary. To assess tone, we collect the number of positive words ($\#pw$), and the number of negative words ($\#nw$), so that, given a sentence s, we can simply define its tone as:

$$tone(s) = \frac{\#pw - \#nw}{\#pw + \#nw}, \text{ with } \; tone(s) \in [-1, 1]$$

Subsequently, the type of the relation (either attack or support) between couples of arguments, and its weight (negative or positive), is determined by analyzing the tone of each argument. Finally, let us observe that, in this phase, there is no need of splitting sentences into tokens, since both sentiment dictionaries filter out stop-words.

3.2. Bipolar Weighted Graph Instantiation

The BWAF instantiation task from the Q&A section of an ECC can be divided into two steps:

1. definition of abstract arguments, and
2. definition of relations between them.

Algorithm 1 BWAF graph edges building

Require: $G = \langle \mathcal{A}, \hat{\mathcal{R}}, w_{\hat{\mathcal{R}}} \rangle$: BWAF graph; $\mathcal{A} = Q \cup A$ where $Q = \{q_i\}$: questions of analysts, $A = \{a_j\}$: answers of executives; $i, j = 1, \ldots, |\mathcal{A}|$.

```
for all q_i ∈ Q do
  tone(q_i) = mean(SC(q_i), LM(q_i))
  for all a_j ∈ A do
    if i < j then
      if tone(q_i) ≤ 0 then
        add an attack ⟨q_i, a_j⟩ in R̂ with weight w_R̂ = tone(q_i)
      else
        add a support ⟨q_i, a_j⟩ in R̂ with weight w_R̂ = tone(q_i)
      end if
      tone(a_j) = mean(SC(a_j), LM(a_j))
      if tone(a_j) ≤ 0 then
        add an attack ⟨a_j, q_i⟩ in R̂ with weight w_R̂ = tone(a_j)
      else
        add a support ⟨a_j, q_i⟩ in R̂ with weight w_R̂ = tone(a_j)
      end if
    end if
  end for
end for
return G
```

Step (1) splits the Q&A part of the ECC transcript into arguments, each of which is associated to a participant of the conference call. To the extent of representing the rightful flow of arguments and counterarguments in the exchange of questions and answers, the set of arguments is partitioned into two subsets: arguments put forward by analysts (i.e., questions), and those put forward by executives (i.e., answers). The arguments are gathered from the ECC transcript considering each paragraph as a single abstract argument. In this step, interventions of the operator, who is in charge of managing the discussion, are neglected. Once all the arguments are collected, we have to connect them through a weighted relation of attack or support. For a given argument, one can infer its sentiment, which is typically described as the degree to which the argument reflects positively or negatively to the company.

Step (2) relates arguments to one another according to a specific criterion. In the Q&A part of an ECC, this task is not trivial as for each question one or more answers may follow. These interactions result in an attack or a support between the question and an answer, and *vice versa*. Since questions by different analysts never refer to previous questions, there is no relation between questions, or between the next analyst's question and the previous answers. Also, executives' answers are not related to each other, since they respond specifically to a question and cannot bridge to the next question by asking them a question.

The algorithm for edge building in BWAF instantiation is provided in Algorithm 1, and can be summarised as follows:

- for each analyst's question q_i, add an attack/support relation starting from it towards all the answers a_j before the next question q_{i+1};
- for each executive's answer a_j, add an attack/support relation starting from it towards the question q_i before the next question q_{i+1}.

The resulting instantiated BWAF has then a particular structure: since questions do not relate to each other, and neither do answers, the graph structure is bipartite, and it is composed by a collection of all complete sub-graphs representing the exchange of arguments between the analyst' questions and the executives' answers in response to it. As an example Figure 3 depicts the BWAF instantiated from the Q&A section of the ECC of Microsoft Corp. in the third quarter of 2012. It is important to note that when the tone of an argu-

ment is totally neutral, and hence equal to 0, the assumption is to assign an attack of strength null.

3.3. Evaluation by BWAF Laplacian Semantics

Once the BWAF has been instantiated, we exploit the Laplacian ranking-based semantics introduced in Section 2. It should be stressed that such a ranking-based semantics is particularly suited to the analysis of bipartite BWAFs[5] instantiated from ECCs given that their structure consists of various fully connected sub-graphs, and given that ECC transcript may be large, the fact that the Laplacian semantics builds on established and computationally well-behaved[6] techniques from matrix algebra make it a good fit for our purposes in this work. We are interested in evaluating the analysts' confidence in the trend of the company, based on the ECC. Therefore, from the Laplacian ranking of arguments in the Q&A, we initially filter only the accepted arguments, i.e., those ones with a ranking greater than or equal to 0. Then, we establish a ranking of *winning questions*, i.e. the ranking of accepted analysts' questions only. We focus on winning questions because questions (and how they are replied to) is what would sway an analyst's opinion.

Let us illustrate the above process through an example:

Example 4. *The BWAF G_2 in Figure 3 yields the following Laplacian-ranking semantics:*
$a3 \succeq^{\deg}_{G_2} q5 \succ^{\deg}_{G_2} q42 \succ^{\deg}_{G_2} q11 \succ^{\deg}_{G_2} q45 \succ^{\deg}_{G_2} a29 \succ^{\deg}_{G_2} a6 \succ^{\deg}_{G_2} a26 \succ^{\deg}_{G_2} a51 \succ^{\deg}_{G_2} a34$, *where*
$\deg(a3) = 0.75$, $\deg(q5) = 0.75$, $\deg(q42) = 0.56$, $\deg(q11) = 0.47$, $\deg(q45) = 0.37$, $\deg(a29) = 0.33$, $\deg(a6) = 0.32$, $\deg(a26) = 0.17$, $\deg(a51) = 0.15$, $\deg(a34) = 0.08$.
We can then identify the winning questions: $W_Q = \{q5, q42, q11, q45\}$, *which are, specifically, the following ones:*

$q5$ = *"I was just wondering, Peter, a few years ago, and you might not have been CFO at the time, you guys had talked about online and what your goals were 5 years out, and talking about 20% organic market share and that could get you to breakeven. I mean, given what we're seeing in terms of RPS, although you're doing a very good job on the OpEx side, how would you say you're thinking about that today?"*
$q42$ = *"I was having a forward-looking question on the gross margin. Remember in the last few quarters where you had a slight negative mix effect there. As I look into the new product launches into the next year, is there anything – they're all kind of high gross margin areas. Is there anything that stops that feeding through in the P&L? Or should it be a straightforward one?"*
$q11$ = *"And will Skype be a big benefit to that division going forward?"*
$q45$ = *"This past quarter, both Gartner and IDC saw a better-than-expected PC uptick in the European corporate market. And your reported Windows revenues certainly support that. I know broadly speaking the business refresh was healthy. Just was wondering if there's anything you would add that might have contributed to an uptick in Europe PC growth."*

3.4. Tone-based Evaluation

For a given Q&A section of an ECC, we need to determine a *relevance value* in order to predict analysts recommendations with a significant positive or negative tone. The working hypothesis is that analysts must be updating their beliefs using argumentative information obtained during these calls. For this reason, [36] studied how analysts revise their beliefs in response to new information depending on the tone of the ECC. Starting from this assumption, we combine tone-based textual analysis and the solution inferred through the BWAF semantics in order to generate a relevance value.

We therefore aggregate Laplacian-based acceptability degrees (Definition 3), which are determined by exploiting the tone of each argument, among the selected ones and from them we determine a final scoring value. For this task, we devise three different scoring functions:

- *Global Average Tone* represents the average tone of the whole Q&A without distinguish-

[5] A further advantage of exploiting the Laplacian matrix of BWAFs for ECCs is that the Laplacian matrix has only nonnegative eigenvalues (it is positive-semidefinite), and its eigenvectors can be used for grouping the nodes of the graph into clusters, and hence enhanced analysis may be run on clusters of questions, or answers, or even better, on sentences from a particular analyst or executive.

[6] Notice in particular that the diameter of the instantiated graphs is always 2.

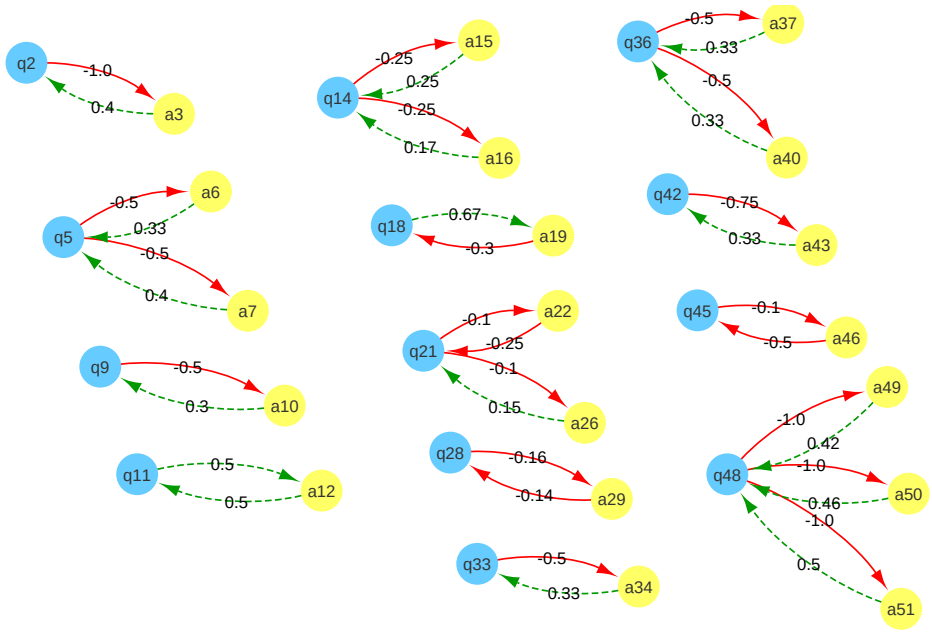

Fig. 3. G_2: BWAF representation of the ECC Q&A part of Microsoft Corp. in the 3rd quarter of 2012.

ing between executives and analysts. Intuitively, investors may simply follow managers' tone in financial disclosures, even though their tone may not exactly represent the underlying fundamentals of the firm. Formally, $\forall \alpha_i \in A$ *s.t.* $|A| = n, i = 1, \ldots, n$:

$$gt(\alpha_i) = \frac{1}{n} \sum_{i}^{n} tone(\alpha_i).$$

– *Analysts Majority Tone* represents the average tone of *winning questions* rankings. Formally, $\forall q_i \in Q$ *s.t.* $|Q| = k$ and $\deg(q_i) > 0, i = 1, \ldots, k$:

$$at(q_i) = \frac{1}{k} \sum_{i}^{k} tone(q_i).$$

– *Weighted Analysts Majority Tone* represents the average tone of *winning questions* rankings, mediated by the number of answers each question receives. Formally, $\forall q_i \in Q, \forall a_j \in A$ *s.t.* $q_i \hat{R} a_j$, $|Q| = k$ and $\deg(q_i) > 0, i = 1, \ldots, k,\ j = 1, \ldots, m$:

$$wat(q_i) = \frac{1}{mk} \sum_{i}^{k} tone(q_i).$$

Example 5. *Continuing the Example 4, we have the following scoring functions:*

– *Global Average Tone:* -0.0297
– *Analysts Majority Tone:* 0.266
– *Weighted Analysts Majority Tone:* 0.011784

4. Experiments

We now study to what extent the framework detailed in the previous section can help predicting analysts' recommendation in buying, holding or selling company's stocks. In what follows, we describe the procedures adopted for:

– gathering the data,
– executing the general processing framework of the NLP sentiment model, BWAF instantiation, Laplacian-ranking semantics evaluation, and final tone-based scoring value,
– learning the machine learning classification model.

4.1. Dataset Construction and Framework Processing

Given the novelty of the study, we needed to build an original dataset. We gathered first the

data on historical analysts' recommendations. Zacks encompasses the full range of investment information required to effectively manage individual and institutional US equity investment processes. Zacks Data[7] can be used to empirically analyze analysts' forecasts and their revisions, price targets and recommendations. This is a proprietary data set, whose historical analyst recommendations data we could access through a free trial for the Wharton Research Data Services (WRDS)[8], which is a data research platform providing access to U.S. equity investment data, market data systems and data from Zacks. We gathered analysts recommendations from 10 companies, from 2007 to 2012. Table 1 reports the list of companies involved in the experiments and their corresponding sector.

The data required to retrieve ECC transcripts came from Seeking Alpha[9], a well-known platform for investment research, with broad coverage of stocks, asset classes, ETFs and investment strategy. This website contains publicly available conference call transcripts for US stocks and ADRs (American Depositary Receipt). We can have free access to the texts online on Seeking Alpha. As there are so many transcripts, getting them manually is very inefficient. With the help of web scraping techniques in Python, and regular expressions, we captured all transcripts automatically. Three main Python libraries were used to scrape the data: BeautifulSoup4, Urllib2 and Requests.

An ECC transcript on Seeking Alpha is made up of four parts:

1. list of executives (E);
2. list of analysts (A);
3. Corporate Presentation Session (CP);
4. Question & Answers Session (Q&A).

Each conference call transcript was then split into three parts, neglecting information about CP, since it is focused only on the message transmitted by the executives team, with no analysts' participation. The Q&A Session part was cleaned by the operator's interventions, and we assigned to each argument, i.e., paragraph in the transcript, the corresponding participant (either an analyst or an executive). Sometimes transcripts may report an unidentified analyst, so we adjusted the assignment to a generic participant qualified as analyst. API for SC and LM dictionary were available in Python and exploited to assess tone, and NetworkX was used to build the BWAF. Then, Laplacian-ranking semantics was developed with Numpy and Scipy. Finally, the three tone-based evaluations were assessed for each ECC transcript. Not all the transcripts were available on Seeking Alpha, especially the oldest ones. Then, by collecting all the available data from WRDS, Seeking Alpha, and Yahoo Finance (for contextual information about companies), the dataset was finally ready for the prediction analysis. The gathered data consisted of 153 entries, and is the first dataset of this kind for financial analysis ever built[10].

4.2. Predicting Analysts Recommendations

In order to predict analysts recommendations, the next step was to generate relevant features. The features are really important because these are what we were suggesting is predictive of the target variable. Our target variable is the recommendation, i.e. a class of the following type:

1. Strong buy;
2. Buy;
3. Hold;
4. Sell;
5. Strong Sell.

In this phase we compare our performance results with a baseline. For this task, our baseline is the overall tone of the whole ECC, i.e. the average sentiment coming from the analysis of the entire transcript, considering both CP and Q&A sessions. For the baseline, this sentiment score associated to each ECC will be the only feature to train our model, since this is the approach currently considered state-of-the-art in financial research on ECCs [37]. Instead, for our argumentation-based approach, the features of our model are the three tone-base evaluation scores, aiming at proving that the underlying rationale of argumentation can better explain the informational relevance of the Q&A part of an ECC.

[7] http://www.zacksdata.com/
[8] https://wrds-web.wharton.upenn.edu/wrds/
[9] https://seekingalpha.com/

[10] The dataset, together with all the scraped ECC transcripts, is available at: https://figshare.com/projects/Earnings_Conference_Calls_Dataset/31370

Table 1
Companies in Stock Market NYSE and NASDAQ

Code	Company	Sector
ALL	Allstaste Corp.	Financial
CBG	CBRE Group, Inc.	Financial
DVN	Devon Energy Corp.	Basic Materials
IBM	IBM Corp.	Technology
MRO	Marathon Oil Corp.	Basic Materials
MSFT	Microsoft Corp.	Technology
ROK	Rockwell Automation, Inc.	Industrial Goods
S	Sprint Corp.	Technology
MTOR	Meritor, Inc.	Consumer Goods
GME	GameStop Corp.	Services

No recommendation for class 5 (Strong Sell) were present in the dataset. Therefore, we dealt with a Machine Learning problem of multi-class classification with 4 classes to be predicted. Since there is not perfect machine learning algorithm for a particular application, we decided to test several machine learning algorithms before a particular algorithm is selected. This is done mainly for the following reasons:

- Evaluate the prediction performance differences between the baseline and our approach;
- Evaluate which machine learning algorithm better fits this kind of financial and sentiment data;
- Discuss on which "argumentative features" and machine learning algorithm may have a preferential choice and a higher impact when facing with the inferential task of using abstract argumentation to classify an object.

Therefore, we chose to run the following machine learning algorithms:

- Generalised Linear Models:
  - Logistic Regression Classifier (LR) [38]
  - Ridge Classifier (RC) [39]
- Support Vector Machine (SVM) [40]
- K-Nearest Neighbors (KNN) [41]
- Gaussian Process Classification (GPC) [42]
- Naive Bayes (NB) [43]
- Decision Tree (DT) [44]
- Ensemble Methods:
  - Random Forest (RF) [45]
  - Gradient Tree Boosting (GTB) [46]
- Neural Networks Model: Multi-layer Perceptron (MLP) [47]

Python libraries Pandas and scikit-learn [48] were exploited for this task. The data was randomly split into testing (20%) and training (80%) sets, and each model was trained and tested. The performance measure to validate the test set was accuracy. In multi-class classification, this corresponds to subset accuracy, which is a harsh metric since it is required for each sample that each label set is correctly predicted. To avoid overfitting, we performed also a 5-fold cross validation. It is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set. The training set was split into 5 smaller sets. A model is trained using 4 of the folds as training data and the resulting model is validated on the remaining part of the data (i.e., it is used as a test set to compute the accuracy performance measure). Then, the activity of splitting the data, fitting the model and computing the score is repeated 5 consecutive times (with different splits each time). The performance measure reported by 5-fold cross-validation is then the average of the values computed in the loop. We hence collect the overall accuracy score, the accuracy of each fold of the cross validation and the related mean score together with the confidence interval of the score estimate.

We report in Table 2 the results obtained for all the machine learning algorithms with the baseline dataset. While in Table 3 are reported the results obtained for all the machine learning algorithms with our argumentative features dataset. We highlight in Table 4 the comparison between our approach and the baseline performances. For all the tables, entries are ordered by decreasing mean accuracy score, giving an immediate overview of

which algorithms achieved better performances (and which ones performed worse).

Regarding the Baseline performances, we note that SVM, MLP, LR, GPC, and RC achieved all the same results, with a mean score accuracy of 51.68%. NB, RF, DT, GBC, and KNN (with K = 3) performed worse, instead. Actually, all baseline performances show bad accuracy scores, thus showing to be not highly predictive. Taking into account our argumentation-based approach, we note that SVM, MLP, and RF achieved higher performances, with a mean score accuracy of 77.27%. LR, GPC, and RC performed quite the same, with $76,6\%$ mean score accuracy. KNN (with K = 3), GBC, DT performed a bit worse, while NB achieved the worst performances. The motivation behind the achievement of the exactly same result for most classifiers lies in the composition, sampling number, and splitting methods of the dataset, since training machine learning models with only 122 training data (with $80-20\%$ train-test splitting) are sometimes not sufficient to fit a good generalised model. Anyway, by looking in particular to Table 4, 9 machine learning classifiers out of 10 performed better with our argumentation-based approach. This gives to our approach a clear value, certifying that machine learning algorithms perform better when argumentation-augmented features have been exploited.

Another insight discovered from our experimental approach regards the choice of a particular machine learning algorithm with better classification performances. We note that SVM and MLP achieved better accuracy scores in both baseline and argumentation-based approaches. In general, the "No Free Lunch" theorem applied to classifiers says that there is no classifier above all [49]. This means that one could always find a case where a classifier is beaten by another. In other words, it is not guaranteed that a particular classifier will perform better than all others. This is the main assumption that encouraged us to run several machine learning algorithms. As general rules of thumb about what to expect from the outcomes we have that, on the one hand, since SVM is obtained by minimizing the structural risk, it is expected to do better than other classifiers. On the other hand, since MLP has the ability to discover the non-linear relationship in the input data set without *a priori* assumption of knowledge of relation between the input and the output, it is expected to achieve good performances, in particular with financial data, given that the existence of the non-linearity and volatility is propounded by many financial analysts. Because of the nature of our fresh dataset, the obtained results may therefore witness that on tone-based financial data, SVM and MLP achieve better performances. This may give a hint for data scientists when facing with tone-based financial data.

5. Conclusions

The paper reported on an application of computational argumentation techniques to the analysis of an important form of financial communication: earnings conference calls (ECCs). Our approach shows that incorporating suitably processed argumentative information in the analysis of ECCs leads to strong predictions of analysts' recommendation, suggesting that argumentative and dialogical features present in ECCs carry informational value for analysts. In doing this we also contributed a novel data set incorporating both argumentative and financial features, as well as a fresh ranking-based semantics for BWAF based on insights from matrix algebra. We put in evidence that computational argumentation can help to improve performances in a classification task due to the fact that the reasoning over conflicting information (which in this case are features of a predictive task) strengthen the informational power of the starting features.

This work represents a first step towards the deployment of computational argumentation techniques in the domain of financial communication. The model built is likely to be improved by including data about more companies and over a longer period of time. We plan to build a wider dataset, that we previously could not do but that our work shows is worth doing. Furthermore, our argumentation-based approach, which focused on analysts' recommendations, may be tested against other forms of financial estimations, such as Earnings Per Share (EPS), Surprise and Estimates prediction, stock returns, and stock prices.

Table 2
Baseline Performances

Machine Learning Algorithm	**Overall Accuracy (%)**	**5-fold Cross Validation Accuracy (%)**						
		Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**CV Mean**	**Confidence Interval**
Support Vector Machine	54.84	50.00	51.61	51.61	51.61	53.57	**51.68**	2.27
Multi Layer Perceptron	54.84	50.00	51.61	51.61	51.61	53.57	**51.68**	2.27
Logistic Regression	54.84	50.00	51.61	51.61	51.61	53.57	**51.68**	2.27
Gaussian Process Classifier	54.84	50.00	51.61	51.61	51.61	53.57	**51.68**	2.27
Ridge Classifier	54.84	50.00	51.61	51.61	51.61	53.57	**51.68**	2.27
Naive Bayes	54.84	50.00	51.61	48.39	51.61	50.00	**50.32**	2.41
Random Forest	51.61	50.00	51.61	48.39	32.26	46.43	**45.74**	13.91
Decision Tree	48.39	50.00	48.39	54.84	35.48	35.71	**44.88**	15.75
Gradient Boosting Classifier	54.84	46.88	41.94	41.94	38.71	35.71	**41.03**	7.45
K-Nearest Neighbors	41.94	37.50	25.81	45.16	38.71	46.43	**38.72**	14.67

Table 3
Argumentation-based Performances

Machine Learning Algorithm	**Overall Accuracy (%)**	**5-fold Cross Validation Accuracy (%)**						
		Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**CV Mean**	**Confidence Interval**
Support Vector Machine	83.87	72.73	75.00	80.00	79.31	79.31	**77.27**	5.77
Multi Layer Perceptron	83.87	72.73	75.00	80.00	79.31	79.31	**77.27**	5.77
Random Forest	83.87	72.73	75.00	80.00	79.31	79.31	**77.27**	5.77
Logistic Regression	83.87	72.73	75.00	76.67	79.31	79.31	**76.60**	5.08
Gaussian Process Classifier	83.87	72.73	75.00	76.67	79.31	79.31	**76.60**	5.08
Ridge Classifier	83.87	72.73	75.00	76.67	79.31	79.31	**76.60**	5.08
K-Nearest Neighbors	67.74	51.52	68.75	73.33	79.31	75.86	**69.75**	19.50
Gradient Boosting Classifier	74.19	57.58	59.38	73.33	62.07	58.62	**62.19**	11.53
Decision Tree	61.29	51.52	56.25	66.67	68.97	65.52	**61.78**	13.43
Naive Bayes	19.35	15.15	18.75	80.00	13.79	13.79	**28.30**	51.83

Table 4
Cross Validation Mean Accuracy Comparison

Machine Learning Algorithm	**Arg.-based**	**Baseline**
Support Vector Machine	77.27	51.68
Multi Layer Perceptron	77.27	51.68
Random Forest	77.27	45.74
Logistic Regression	76.60	51.68
Gaussian Process Classifier	76.60	51.68
Ridge Classifier	76.60	51.68
K-Nearest Neighbors	69.75	38.72
Gradient Boosting Classifier	62.19	41.03
Decision Tree	61.78	44.88
Naive Bayes	28.30	50.32

References

[1] B. Crawford Camiciottoli, *Rhetoric in financial discourse. A linguistic analysis of ICT-mediated disclosure genres*, Rodophi, 2013.

[2] R. Palmieri, A. Rocci and N. Kudrautsava, Argumentation in earnings conference calls. Corporate standpoints and analysts' challenges, *Studies in Communication Sciences* **15**(1) (2015), 120–132.

[3] K. Budzynska, A. Rocci and O. Yaskorska, Financial Dialogue Games: A Protocol for Earnings Conference Calls, in: *Computational Models of Argument - Proceedings of COMMA 2014*, 2014, pp. 19–30.

[4] A. Rocci and C. Raimondo, Conference calls: a communication perspective, in: *Handbook of investor relations and financial communications*, A. Laskin, ed., Wiley & Sons, 2017.

[5] D. Matsumoto, M. Pronk and E. Roelofsen, What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions, *The Accounting Review* **86**(4) (2011), 1383–1414.

[6] M. Lippi and P. Torroni, Argumentation Mining: State of the Art and Emerging Trends, *ACM Transactions on Internet Technology* **16**(2) (2016), 10–11025, ISSN 1533-5399.

[7] S. Menini, E. Cabrio, S. Tonelli and S. Villata, Never Retreat, Never Retract: Argumentation Analysis for Political Speeches, in: *Proceedings of AAAI 2018*, 2018.

[8] O. Cocarascu and F. Toni, Identifying attack and support argumentative relations using deep learning, in: *Proceedings of EMNLP 2017*, 2017, pp. 1374–1379.

[9] V. Niculae, J. Park and C. Cardie, Argument Mining with Structured SVMs and RNNs, in: *Proceedings of 55th Annual Meeting of the ACL*, 2017, pp. 985–995.

[10] E. Cabrio and S. Villata, Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions, in: *Proceedings of 50th Annual Meeting of the ACL*, 2012, pp. 208–212.

[11] S. Rosenthal and K. McKeown, Detecting Opinionated Claims in Online Discussions, in: *Proceedings of 6th International Conference of Semantic Computing*, ICSC '12, IEEE Computer Society, 2012, pp. 30–37. ISBN ISBN 978-0-7695-4859-3.

[12] A. Rosenfeld and S. Kraus, Providing arguments in discussions on the basis of the prediction of human argumentative behavior, *ACM Transactions on Interactive Intelligent Systems* **6**(4) (2016), 30.

[13] I. Habernal and I. Gurevych, What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation, in: *Proceedings of EMNLP 2016*, 2016, pp. 1214–1223.

[14] I. Habernal and I. Gurevych, Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse, in: *Proceedings of EMNLP 2015*, 2015, pp. 2127–2137.

[15] I. Habernal and I. Gurevych, Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM, in: *Proceedings of 54th Annual Meeting of the ACL*, 2016, pp. 1589–1599.

[16] R. Kaptein, M. Marx and J. Kamps, Who said what to whom?: capturing the structure of debates, in: *Proceedings of the 32nd Annual International ACM SIGIR 2009*, 2009, pp. 831–832.

[17] Z. Salah, F. Coenen and D. Grossi, Extracting debate graphs from parliamentary transcripts: a study directed at UK house of commons debates, in: *Proceedings of the 14th International Conference on Artificial Intelligence and Law*, ICAIL '13, 2013, pp. 121–130.

[18] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G.R. Simari, T. M. and S. Villata, Towards Artificial Argumentation, *AI Magazine* **38**(3) (2017), 25–36.

[19] S. Polberg and A. Hunter, Empirical Evaluation of Abstract Argumentation: Supporting the Need for Bipolar and Probabilistic Approaches, *International Journal of Approximate Reasoning* **93** (2018), 487–543.

[20] A. Pazienza, S. Ferilli and F. Esposito, Constructing and Evaluating Bipolar Weighted Argumentation Frameworks for Online Debating Systems, in: *Proceedings of the 1st Workshop on Advances In Argumentation In Artificial Intelligence, co-located with XVII International Conference of the Italian Association for Artificial Intelligence, AI^3@AI*IA 2017*, 2017, pp. 111–125.

[21] S.M. Price, J.S. Doran, D.R. Peterson and B.A. Bliss, Earnings conference calls and stock returns: The incremental informativeness of textual tone, *Journal of Banking & Finance* **36**(4) (2012), 992–1011, ISSN 0378-4266.

[22] H. Wachsmuth, J. Kiesel and B. Stein, Sentiment Flow - A General Model of Web Review Argumentation, in: *Proceedings of EMNLP 2015*, 2015, pp. 601–611.

[23] P.M. Dung, On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games, *Artificial Intelligence* **77**(2) (1995), 321–357.

[24] G.R. Simari and I. Rahwan (eds), *Argumentation in Artificial Intelligence*, Springer, 2009.

[25] P. Baroni, D. Gabbay, M. Giacomin and L. van der Torre, *Handbook of Formal Argumentation*, Vol. 1, College Publications, 2018. ISBN ISBN 9781848902756.

[26] C. Cayrol and M. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: *Proceedings of ECSQARU 2005*, Springer, 2005, pp. 378–389.

[27] P.E. Dunne, A. Hunter, P. McBurney, S. Parsons and M. Wooldridge, Weighted Argument Systems: Basic Definitions, Algorithms, and Complexity Results, *Artificial Intelligence* **175**(2) (2011), 457–486, ISSN 0004-3702.

[28] A. Pazienza, S. Ferilli and F. Esposito, On the Gradual Acceptability of Arguments in Bipolar Weighted Argumentation Frameworks with Degrees of Trust, in: *Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017*, 2017, pp. 195–204.

[29] L. Amgoud and J. Ben-Naim, Ranking-based semantics for argumentation frameworks, in: *International Conference on Scalable Uncertainty Manage-*

ment, Springer, 2013, pp. 134–147.

[30] R.W. Floyd, Algorithm 97: shortest path, *Communications of the ACM* **5**(6) (1962), 345.

[31] A. Pazienza and S. Ferilli, The Linear Algebra of Abstract Argumentation, in: *Proceedings of the 2nd Workshop on Advances In Argumentation In Artificial Intelligence, co-located with XVII International Conference of the Italian Association for Artificial Intelligence, AI^3@AI*IA 2018*, 2018, pp. 71–85.

[32] J.M. Ortega, *Matrix Theory: A Second Course*, University Series in Mathematics, Springer US, 2013. ISBN ISBN 9781489904713.

[33] F.R. Gantmakher, *The theory of matrices*, Vol. 1, Chelsea Publishing Company, New York, 1959, Chap. 5, Section 4, Theorem 1.

[34] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard and D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: *ACL (System Demonstrations)*, 2014, pp. 55–60.

[35] T. Loughran and B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance* **66**(1) (2011), 35–65.

[36] C. Chen, C. Liu, Y. Chang and H. Tsai, Opinion mining for relating subjective expressions and annual earnings in US financial statements, *Journal of Information Science and Engineering* **29**(2) (2012).

[37] P.A. Borochin, J.E. Cicon, R.J. DeLisle and S.M. Price, The effects of conference call tones on market perceptions of value uncertainty, *Journal of Financial Markets* (2018), ISSN 1386-4181.

[38] D.W. Hosmer Jr, S. Lemeshow and R.X. Sturdivant, *Applied logistic regression*, Vol. 398, John Wiley & Sons, 2013.

[39] A.E. Hoerl and R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1) (1970), 55–67.

[40] C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning* **20**(3) (1995), 273–297.

[41] T. Cover and P. Hart, Nearest Neighbor Pattern Classification, *IEEE Trans. Inf. Theor.* **13**(1) (1967), 21–27, ISSN 0018-9448.

[42] C.E. Rasmussen, Gaussian processes in machine learning, in: *Advanced lectures on machine learning*, Springer, 2004, pp. 63–71.

[43] H. Zhang, The optimality of naive Bayes, *AA* **1**(2) (2004), 3.

[44] L. Breiman, *Classification and regression trees*, Routledge, 2017.

[45] L. Breiman, Random forests, *Machine learning* **45**(1) (2001), 5–32.

[46] J.H. Friedman, Greedy function approximation: a gradient boosting machine, *Annals of statistics* (2001), 1189–1232.

[47] G.E. Hinton, Connectionist learning procedures, in: *Machine Learning, Volume III*, Elsevier, 1990, pp. 555–610.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825–2830.

[49] D.H. Wolpert and W.G. Macready, No free lunch theorems for optimization, *IEEE transactions on evolutionary computation* **1**(1) (1997), 67–82.