A Borda Count for Collective Sentiment Analysis

Umberto Grandi[*] Andrea Loreggia[†] Francesca Rossi[‡]

Vijay Saraswat[§]

To appear on *Annals of Mathematics and Artificial Intelligence*, special issue on ”Preferences and Computational Social Choice”, 2015

Abstract

Sentiment analysis assigns a positive, negative or neutral polarity to an item or entity, extracting and aggregating individual opinions from their textual expressions by means of natural language processing tools. In this paper we observe that current sentiment analysis techniques are satisfactory in case there is a single entity under consideration, but can lead to inaccurate or wrong results when dealing with a set of multiple items. We argue in favor of importing techniques from voting theory and preference aggregation to provide a more accurate definition of the collective sentiment over a set of multiple items. We propose a notion of Borda count which combines individuals' sentiment with comparative preference information, we show that this class of rules satisfies a number of properties which have a natural interpretation in the sentiment analysis domain, and we evaluate its behavior when faced with highly incomplete domains.

1 Introduction

We live in a world where we communicate more and more on social media, writing text that reflects our opinions and feelings. Being able to formalize such opinions and reason with them can be very useful for a number of practical applications. First, service providers may personalize their offer based on customers opinions. Second, companies may test what products would be better received by potential consumers, and adjust their strategy accordingly. Third, community councils and candidates in political elections may evaluate the reception of their proposals, and focus their attention on the most preferred ones. It comes therefore as no surprise that the extraction of individual opinions from textual expressions, such as tweets, blog posts, or product reviews, has been the subject of a very active area of research in recent years.

[*] Department of Mathematics, University of Padova. Current affiliation: IRIT, University of Toulouse. Corresponding author at `umberto.grandi@ut-capitole.fr`

[†] Department of Mathematics, University of Padova.

[‡] Department of Mathematics, University of Padova. Current affiliation: IBM TJ Watson Research Lab, Yorktown Heights, USA.

[§] IBM TJ Watson Research Lab, Yorktown Heights, USA.

Researchers in sentiment analysis and opinion mining (34; 40) developed a collection of tools in natural language processing (NLP) for the extraction of opinions, sentiments, or attitudes of individuals from their textual expressions. In order to summarize the opinion of all the individuals in a unique indicator, the opinions extracted are then used to define a notion of collective sentiment about the entities under consideration, be they commercial products, policies or candidates.

In this paper we observe that current sentiment analysis techniques are good enough when we are trying to understand the positive or negative opinion of a set of agents over a single item, but they fall short when we are considering several items. Our claim stems from the observation that, when several items are being compared, the approach taken by sentiment analysis of only focusing on positive or neutral polarities may differ from the approach that computes the most preferred item by making use of comparative preference information. Consider for instance the following situation, in which two candidates Ann and Bob are competing in an election.

Example 1. *Assume there are a total of 35 people who are expressing their positive or negative attitude on social media about two candidates Ann and Bob: 20 persons are talking positively about Ann, 15 persons are talking negatively about Ann, 30 persons are talking positively about Bob, and 5 persons are talking negatively about Bob. However, what people write on social media is just a textual abstraction of the comparative preferences they have in mind, which in this particular case we assume to be a ranked list of the two candidates. Assume therefore that their preferences are as described in the following table, where candidates to the left are more preferred than candidates to the right, and the bar signals the threshold of positive vs. negative opinions:*

20 voters:	Ann	Bob	|		
10 voters:		Bob	|	Ann	
5 voters:			|	Ann	Bob
Sentiment analysis				Bob	
Majority rule				Ann	

In the profile described above there are 30 voters that express a positive opinion about Bob, and 20 voters that express a similar opinion about Ann. Hence, sentiment analysis, as well as similar methods based solely on sentiment information, would conclude that Bob is the most popular candidate. However, if we assume that the election will be decided by majority voting (which is the only sensible rule to be used when deciding among two candidates), then Ann will be the winner of the election with 25 votes over 10 for Bob, unlike the outcome of sentiment analysis. Observe that the positive/negative opinions expressed by the individuals are consistent with the preferences that will then be revealed at the time of voting.

The situation above is a good example of the use of sentiment analysis and preference aggregation for the *prediction* of a real-world event. In this particular case, a prediction of an electoral result is being based on the number of positive opinions extracted from voters. Similar examples can also be devised to point out a problem in situations of *decision-making*: think of Ann and Bob as two products that a firm is considering to promote, and the sentiment and preferences expressed in the table be those extracted from conversations and reviews of its customers. When the firm needs to decide which of the two products to invest in, sentiment analysis and preference aggregation would give two different recommendations.

The first message of this paper is that all these considerations can be phrased in the framework of preference analysis (43; 28) and voting theory (4). For instance, sentiment analysis as presented in the example above uses a preference aggregation method called approval voting (11), which is based only on positive or negative opinions expressed over candidates. Text-extracted opinions may present both polarities and preference orderings, and the main contribution of this paper is to propose a definition of collective sentiment that makes use of both kinds of information.

Building on the classical Borda count (see, e.g., (10)) we define and study a class of voting rules that aggregate both polarities and preference orderings into a collective sentiment, taking into account the incompleteness inherent in text-extracted opinions, where each individual may refer only to some of the items under consideration. We study the behavior of this class of rules from a decision-theoretic perspective. First, we list a number of properties that are desirable in the context of sentiment analysis, and we show that our proposed rules satisfy all such conditions. Second, we perform experiments to quantify the discrepancy between classical sentiment analysis techniques and our proposed rule, and we investigate its behavior with respect to partial information. The results we obtain indicate that our proposed Borda count not only satisfies a list of desirable properties when its outcome is used as a basis for decision-making, but also it is computationally tractable and it behaves well in highly incomplete domains.

To the best of our knowledge this paper is the first attempt to apply techniques from preference aggregation and voting theory to sentiment analysis over multiple issues. Related work has focused on sketching a road map for developing sentiment analysis as an alternative to opinion polls for the prediction of electoral results (37), focusing however on the statistical significance of the population studied rather than on the aggregation method used. Preference aggregation techniques have been used with success in other areas of computer science such as human computation and collective annotation of textual corpora (21; 35), and on developing procedures for collective decision making that are able to handle incomplete preferences (42; 44). A line of work which is similar in spirit to the one proposed in this paper is the work of Brams and Sanver (9) in social choice theory, albeit for the specific setting of committee decisions and elections. We refer to Section 4.3 for a more detailed discussion of this approach. We also acknowledge the work of Garg et al. (26) on opinion pooling, which is however focused on the aggregation of probabilistic opinions. The present work expands our previous position paper on the use of preferences and voting techniques in sentiment analysis (27).

The paper is organized as follows. In Section 2 we present the basic concepts and definitions of the framework of sentiment analysis and that of voting theory. In Section 3 we provide a formal definition of preferential information extracted from text, we formalize the two classical approaches of sentiment analysis and voting theory, and we present our novel definition of preference structure that combines sentiment with comparative preference. In Section 4 we put forward our definition of a Borda count for collective sentiment analysis and we evaluate rules from this class from both an axiomatic and an algorithmic perspective. Section 5 presents an experimental analysis of the differences between classical sentiment analysis and our proposed Borda count, and evaluates it in conditions of data sparsity. Section 6 concludes the paper and points at a number of challenges for future work in applying techniques from preference aggregation and voting theory to the domain of sentiment analysis.

2 Background

In this section we present the basic definitions of sentiment analysis, voting theory and preference aggregation.

2.1 Sentiment Analysis

Sentiment analysis and opinion mining (40; 34) is a collection of techniques for the extraction of people's opinions, sentiments, and evaluations from textual expressions. A set of entities or alternatives $\mathcal{X}$ is defined as the sentiment targets, and individual opinions about entities in $\mathcal{X}$ are extracted from a given set of product reviews, blog posts or other sources of textual information.

Formally, two forms of opinions can be identified:

Definition 1. *(33) A* regular opinion *is a tuple (g, s, h, t) where g is the sentiment target, s is the sentiment about the target, h is the opinion holder and t is the opinion time.*

Definition 2. *(30) A* comparative opinion $(e_1, e_2, \mathrm{pa}, h, t)$ *is a tuple where e_1 and e_2 are two entities that are being compared,* pa *is the preferred alternative among e_1 and e_2, h is the opinion holder and t the time.*

Sentiment targets are also called entities or items, and can be anything such as products, policies or persons. The sentiment s in a regular opinion is usually taken to be a positive, negative or neutral polarity, i.e., an element of $\{+, -, 0\}$, although recent developments are directed to a more general setting of graded polarity such as a "five-stars" scale or numerical score (22). The opinion holder h is the individual who wrote a text expressing sentiment s, and the time t is the moment at which h wrote the text. In this paper we will not make use of the temporal information, but we refer the reader to Section 6 for further discussion on the important role that temporal information may play in the development of principled notions of collective sentiment.

The objective of sentiment analysis is to extract all possible opinion tuples as in Definitions 1 and 2. Popular approaches to perform this task use a bag of words extracted from a tagged corpus of positive sentences, and then count in a more or less complex way the presence of such positive words in untagged documents (41). Machine learning techniques such as naive Bayes approaches and sentiment classificators built using semi-supervised learning are also widely used for these tasks (see, e.g., (41; 5) for regular opinions and (30; 25) for comparative opinions).

A notion of *collective sentiment* aggregates individuals' opinions into a collective view, and it is usually expressed as a polarity. The most common approaches define the collective sentiment as a positive sentiment if the number of positive opinions about the item outnumbers the number of negative opinions. When more than one item is considered, each textual expression is classified as positive, negative or neutral, and the items with the largest number of positive expressions are declared as the most preferred ones according to the collective opinion (see, e.g., (38; 6; 13)).

2.2 Voting Theory

There is a wide literature in the field of economic theory and, more recently, artificial intelligence, which studies the problem of aggregating the preferences of a set of agents into a common collective choice or preference (4). Formally, the problem is defined by a set of individuals $\mathcal{I}$ expressing preferences over a number of alternatives $\mathcal{X}$, and by a voting rule F which associates a set of winning candidates with such preferences. Preferences can be specified in many different ways, for instance rankings, approvals or a set of binary comparisons. In this paper we consider two such definitions. The first and more common approach represents preferences as *linear orders*, i.e., transitive, anti-symmetric and complete binary relations over $\mathcal{X}$. A profile of preferences $\boldsymbol{P} = (<_1, \ldots, <_n)$ is defined by the choice of a preference relation $<_i$ for each of the n individuals. We write $a <_i b$ to denote that agent i prefers item b to item a in profile $\boldsymbol{P}$. A (non-resolute) *voting rule* F associates with every profile $\boldsymbol{P}$ a non-empty subset of winning candidates $F(\boldsymbol{P}) \in 2^{\mathcal{X}} \setminus \emptyset$. A large number of voting rules have been proposed in the literature (see, e.g., Brams and Fishburn, (10)), and we now present the definition of a widely used procedure that we will later use to ground our definition of collective sentiment:

Definition 3 (Borda rule)**.** *Given a linear order $<_i$ for each $i \in \mathcal{I}$, the* Borda rule *assigns to each alternative $c \in \mathcal{X}$ one point for each alternative that is ranked lower than c in $<_i$, and then takes the sum over all individuals. The alternatives with the highest overall score are elected as the winners.*

The Borda rule is a widely studied voting procedure that can easily be adapted to different preference models, such as orders with ties (23) and partial orders (24; 1; 16). An axiomatic characterization of this rule was first presented by Young (45).

The second approach that we will consider represents individual preferences as a set of approved alternatives. In this case, the only information collected from individuals is whether they approve or not a given alternative, and the following procedure is used to decide the winning alternative:

Definition 4 (Approval voting)**.** *Given a subset of approved alternatives $A_i \subseteq \mathcal{X}$ for each $i \in \mathcal{I}$, the winners of* approval voting *are the candidates that receive the highest number of approvals.*

Despite its simple definition, approval voting has been the subject of an extensive literature since its first appearance (see, e.g., (32)).

Both approval voting and the Borda rule can be adapted to output a ranking of the candidates (from higher to lower score) transforming the two voting rules into *social welfare functions* (43), i.e., functions which associate with every profile of preferences a ranking of the alternatives.

Research in voting theory and preference aggregation has focused on the evaluation of different definitions of collective preference, either by means of axiomatic properties that specify desirable characteristics of the outcome of aggregation, by running experiments on realistic preference distributions, or by analyzing their computational properties such as the complexity of determining the winner (4; 12). We refer to Section 4.1 and 4.4 for a more detailed discussion of these aspects.

3 How to model individuals' opinions

Sentiment analysis and preference aggregation take two different approaches in the representation of absolute and comparative preferential information that is extracted from individual data. The aim of this section is to formally define these two approaches, and to propose a novel structure for preference representation that combines sentiment polarity with comparative preferences.

3.1 Individual data

We assume to have collected a set of textual expressions $\mathcal{T}_i$ for every individual $i \in \mathcal{I}$, and that exploiting tools from NLP we are able to extract regular opinions expressed by individuals about the entities in a set $\mathcal{X}$ in the form of a score (see Definition 1), as well as comparative opinions in the form of binary comparisons (see Definition 2).

Definition 5. *The* individual data *extracted from a set of individual expressions $\mathcal{T}_i$ is a tuple $(\sigma_i, \leqslant_i^P, \leqslant_i^N)$ where:*

- *$\sigma_i : D_i \to \mathbb{R}$ is a function defined on a subset of entities $D_i \subseteq \mathcal{X}$ representing all regular opinions, i.e., degrees of positive and negative opinions over entities;*
- *$\leqslant_i^P$ is a preorder with domain $P_i \subseteq \mathcal{X}$, representing the set of positive comparative opinions of individual i;*
- *$\leqslant_i^N$ is a preorder with domain $N_i \subseteq \mathcal{X}$ representing the set of negative comparative opinions of individual i.*

We make the further assumption that the individual data is always coherent, i.e., the sets P_i and N_i are disjoint sets and when entities a and b are in P_i (resp. N_i) then both $\sigma_i(a)$ and $\sigma_i(b)$ are positive numbers (resp. negative), and also that $a \not\leqslant_i^P b$ (resp. $a \not\leqslant_i^N b$) if $\sigma_i(b) < \sigma_i(a)$. Observe moreover that the sets D_i, P_i and N_i may have non-empty intersection.

Example 2. *A company wants to evaluate three products of different colors: red (R), green (G) and blue (B). A corpora of textual expressions by three individuals is collected and the individual data extracted is as follows. The first individual has a positive opinion about all three products R, G, B, but the degree of these opinions is slightly different: we extract a score of 5 for product R, a score of 4 for entities G and B, and no pairwise comparison among the products. The second individual has a positive score of 1 about product G while expressing a negative opinion about the other two* *colors. She also expresses a direct preference of R over B. Finally, the third individual has a neutral opinion about R and B, while she considers alternative G negatively with a score of -4. We can summarize the opinions extracted from the three individuals in the terminology of our Definition 5:*

- *Individual 1:* $\sigma_1(R) = 5, \sigma_1(G) = \sigma_1(B) = 4$ *and* $P_1 = N_1 = \emptyset$*;*
- *Individual 2:* $\sigma_2(G) = 1$, $P_2 = \emptyset$, *and* $N_2 = \{R, B\}$ *with* $B \leqslant_2^N R$*;*
- *Individual 3:* $\sigma_3(R) = \sigma_3(B) = 0$, $\sigma_3(G) = -4$, *and* $P_3 = N_3 = \emptyset$.

3.2 The sentiment analysis approach

Sentiment analysis (at least in its most common implementation) disregards the intensity of sentiment as well as the comparative opinions, focusing only on the extraction of a positive, negative or neutral polarity from individual expressions.

Definition 6. *Given individual data* $(\sigma_i, \leqslant_i^P, \leqslant_i^N)$ *extracted from individual expressions* $\mathcal{T}_i$*, the* pure sentiment data *associated with it is a function* $\mathit{Sent}_i : E_i \rightarrow \{+, -, 0\}$*, where* $E_i = D_i \cup P_i \cup N_i$*, defined as:*

$$\mathit{Sent}_i(c) = \begin{cases} \mathrm{sgn}(\sigma_i(c)) & \textit{if } c \in D_i \\ 0 & \textit{if } \sigma_i(c) = 0 \\ + & \textit{if } c \in P_i \\ - & \textit{if } c \in N_i \end{cases}$$

After the information about the individual sentiments have been extracted, the most common approach in the definition of the collective sentiment is to choose the entities with the largest amount of positive opinions, disregarding the number of negative opinions.

Example 3. *The pure sentiment data associated with Example 2 is the following:*

- *Individual 1:* $\mathit{Sent}_1(R) = \mathit{Sent}_1(G) = \mathit{Sent}_1(B) = +$
- *Individual 2:* $\mathit{Sent}_2(G) = +$ *and* $\mathit{Sent}_2(R) = \mathit{Sent}_2(B) = -$
- *Individual 3:* $\mathit{Sent}_3(R) = \mathit{Sent}_3(B) = 0$ *and* $\mathit{Sent}_3(G) = -$

Using approval voting as a definition of the collective sentiment we obtain G as the most preferred entity, with two positive opinions received. With the same method we can easily construct a collective ranking of the entities, obtaining R and B tied in the second position with just one positive opinion received.

3.3 The voting theory approach

While sentiment analysis focuses only on polarities, the other extreme of the spectrum is the approach of voting theory, that is restricted to comparative preference information only. For the purpose of this paper we represent individual preferences by using preorders, i.e., reflexive and transitive binary relations. This choice is motivated by two important characteristics of preferences extracted from textual expressions:

- *Interpersonal incomparability*: Since individuals have very different styles of writing or attitudes towards judging the entities under consideration, we believe that scores or any other form of graded polarity cannot be compared across individuals. Therefore we argue in favor of an ordinal representation of both regular and comparative opinions.
- *Incompleteness*: Since preferences and sentiments are observed from individual expressions we cannot assume this information to be complete.

Formally, we can define the voting theory approach as follows:

Definition 7. *Given the individual data $(\sigma_i, \leqslant_i^P, \leqslant_i^N)$ extracted from individual expressions $\mathcal{T}_i$, the* pure preference data *associated with it is a preordered set $(\mathcal{D}_i, \leqslant_i^{\mathcal{D}})$, where $\mathcal{D}_i = D_i \cup P_i \cup N_i$, defined as:*

$$x \leqslant_i^{\mathcal{D}} y \Leftrightarrow \begin{cases} x \leqslant_i^P y \textit{ and } x, y \in P_i & \textit{or} \\ x \leqslant_i^N y \textit{ and } x, y \in N_i & \textit{or} \\ x \in N_i \textit{ and } y \in P_i & \textit{or} \\ \sigma_i(x) \leqslant \sigma_i(y) \textit{ and } x, y \in D_i & \end{cases}$$

Pure preference data is thus the union of the comparative opinions extracted from the ordinal relation entailed by σ_i, with the addition of all the binary comparisons between elements from P_i and elements from N_i.

Example 4. *The pure preference data associated with Example 2 is the following:*

- *Individual 1: $B \sim_1 G <_1 R$*
- *Individual 2: $B <_2 R <_2 G$*
- *Individual 3: $G <_3 B \sim R$*

Where $G < R$ stands for $G \leqslant R$ and $R \nleqslant G$, and $G \sim R$ stands for $G \leqslant R$ and $R \leqslant G$. Using a straightforward adaptation of the Borda rule to preorders, we obtain $B < G < R$ as the collective ranking: R receives 4 points, one for each alternative that is strictly ranked below by one of the individuals, G receives 2 points, and B only 1 point.

3.4 Combining Sentiment with Preference

In this section we propose a novel structure that combines features from both the sentiment analysis approach and the voting theory approach presented in the previous two sections. On the one hand we take binary comparisons as central to our analysis, using preorders to represent comparative preferential information. On the other hand, we complement this representation with a classification of the alternatives into three disjoint sets representing the positive, negative and neutral polarity:

Definition 8. *An* SP-structure (for Sentiment-Preference structure) *over a set of candidates $\mathcal{X}$ is a tuple $S = (\mathcal{P}, \mathcal{N}, \mathcal{Z})$, where $\mathcal{P}$, $\mathcal{N}$ and $\mathcal{Z}$ are disjoint subsets of $\mathcal{X}$, and both $\mathcal{P}$ and $\mathcal{N}$ are ordered respectively by preorders $\leqslant^{\mathcal{P}}$ and $\leqslant^{\mathcal{N}}$.*

An SP-structure indicates the subsets of positive ($\mathcal{P}$), negative ($\mathcal{N}$) and neutral ($\mathcal{Z}$) candidates among the set of entities $\mathcal{X}$, and specifies a set of binary comparisons between positive or negative candidates. The remaining elements of $\mathcal{X} \setminus (\mathcal{P} \cup \mathcal{N} \cup \mathcal{Z})$ are those alternatives for which no information has been collected.

We obtain SP-structures from individual data as follows:

Definition 9. *Let $(\sigma_i, \leqslant_i^P, \leqslant_i^N)$ be the individual data extracted from individual expressions $\mathcal{T}_i$. The SP-structure associated with it is the tuple $(\mathcal{P}_i, \mathcal{N}_i, \mathcal{Z}_i)$:*

- *$\mathcal{P}_i = P_i \cup D_i^+$ where $D_i^+ = \{x \in D_i \mid \sigma_i(x) > 0\}$*
- *$\mathcal{N}_i = N_i \cup D_i^-$ where $D_i^- = \{x \in D_i \mid \sigma_i(x) < 0\}$*
- *$\mathcal{Z}_i = \{x \in D_i \mid \sigma_i(x) = 0\}$*

with preorder relations defined as follows:

$$x \leqslant_i^{\mathcal{P}} y \quad \Leftrightarrow \quad \begin{cases} x \leqslant_i^P y \textit{ and } x, y \in P_i & \textit{or} \\ \sigma_i(x) \leqslant \sigma_i(y) \textit{ and } x, y \in D_i^+ \end{cases}$$

$$x \leqslant_i^{\mathcal{N}} y \Leftrightarrow \begin{cases} x \leqslant_i^N y \textit{ and } x, y \in N_i & \textit{or} \\ \sigma_i(x) \leqslant \sigma_i(y) \textit{ and } x, y \in D_i^- \end{cases}$$

Example 5. *The SP-structures associated with Example 2 are the following:*

- *Individual 1: $\mathcal{P}_1 = \{R, G, B\}$ with $G \leqslant_1^{\mathcal{P}} R$ and $G \sim_1^{\mathcal{P}} B$, $\mathcal{N}_1 = \mathcal{Z}_1 = \emptyset$*
- *Individual 2: $\mathcal{P}_2 = \{G\}$, $\mathcal{N}_2 = \{R, B\}$ with $B \leqslant_2^{\mathcal{N}} R$, $\mathcal{Z}_2 = \emptyset$*
- *Individual 3: $\mathcal{P}_3 = \emptyset$, $\mathcal{N}_3 = \{G\}$ and $\mathcal{Z}_3 = \{R, B\}$*

SP-structures can be easily visualized, and in Figure 1 we draw the three structures described above. Alternatives that are in higher positions in the table are preferred to those that are in lower positions, and the three sets $\mathcal{P}$, $\mathcal{Z}$ and $\mathcal{N}$ are separated by horizontal lines.

Individual 1	Individual 2	Individual 3	
R $\mid$ $G \sim B$	G		$\mathcal{P}$
		R, B	$\mathcal{Z}$
	R $\mid$ B	G	$\mathcal{N}$

Figure 1: SP-structures associated with Example 2.

In conclusion, an SP-structure compactly represents both sentiment information in the form of three polarity sets, as well as comparative opinions in the two preorders over the positive and negative sets. SP-structures are based on a purely ordinal view of preferences, hence assuming a very low degree of interpersonal comparability among individuals' preferences. This assumption could be relaxed by, for instance, normalizing the scores extracted from the individual data, or directly using them in the construction of the collective sentiment. While these approaches may fit some particular applications, they require additional important assumptions when merging comparative opinions, which are of an ordinal nature, with the possibly normalized scores used to represent regular opinions.

4 Borda counts for aggregating SP-structures

In order to aggregate SP-structures, and therefore put forward our definition of collective sentiment, in this section we define a class of aggregation procedures based on the classical Borda count (see Definition 3). We begin by introducing a list of properties which are desirable in the context of sentiment analysis, and then put forward our definition of aggregation method. We show that this method satisfies all the desirable properties we introduced and that it generalizes both the existing definition used by sentiment analysis and the classical Borda rule in preference aggregation. We conclude the section with a study of the algorithmic aspects of our proposed Borda count.

4.1 Desired Axiomatic Properties

The axiomatic method consists in first specifying a number of desirable properties about the problem at hand, and then show a solution that satisfies them, or prove that there exists none. In this section we adapt classical axiomatic properties from the literature in social choice theory to the case of SP-structures, providing a suitable interpretation in the domain of sentiment analysis. We build on the axiomatization of the Borda rule proposed by Young (45), which we complement with axioms specific to our domain of application.

We first need to introduce some useful notation. Let us call a collection of SP-structures $(S_1, \dots, S_n)$ a *profile*, which we denote by $\boldsymbol{S}$. If $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are profiles of SP-structures, let $\boldsymbol{S}_1 + \boldsymbol{S}_2$ be the profile obtained by putting together the two original profiles (renaming voters if necessary), i.e. if $\boldsymbol{S}_1 = (P_1, \dots, P_n)$ and $\boldsymbol{S}_2 = (Q_1, \dots, Q_m)$, then $\boldsymbol{S}_1 + \boldsymbol{S}_2 = (P_1, \dots, P_n, Q_1, \dots, Q_m)$. A profile is called *symmetric* if the set of individuals can be partitioned in pairs of individuals $\{i, i'\}$ with completely opposite SP-structures, i.e., if $\mathcal{P}_i = \mathcal{N}_{i'}$, $\mathcal{N}_i = \mathcal{P}_{i'}$, $\leqslant_i^{\mathcal{P}} = \overleftarrow{\leqslant_{i'}^{\mathcal{N}}}$ and $\leqslant_i^{\mathcal{N}} = \overleftarrow{\leqslant_{i'}^{\mathcal{P}}}$, where $\leqslant_i^{\mathcal{P}} = \overleftarrow{\leqslant_{i'}^{\mathcal{N}}}$ means that the preorder over the set $\mathcal{P}$ for voter i is equal to the inverted preorder over the set $\mathcal{N}$ of voter i'. A symmetric profile necessarly contains an even number of SP-structures. Finally, given a single SP-structure $S = (\mathcal{P}, \mathcal{N}, \mathcal{Z})$, we say that a *voter i ranks a above b* in S if one of the following four conditions holds: $b \leqslant_i^{\mathcal{P}} a$ and $a \not\leqslant_i^{\mathcal{P}} b$, or $b \leqslant_i^{\mathcal{N}} a$ and $a \not\leqslant_i^{\mathcal{N}} b$, or $a \in \mathcal{P}$ and $b \in \mathcal{Z} \cup \mathcal{N}$, or $a \in \mathcal{Z}$ and $b \in \mathcal{N}$.

Let F be a rule which associates a set of most preferred alternatives with a profile of SP-structures. We now list a number of desirable properties for such an aggregation rule F.

The first set of properties is an adaptations of classical axioms from social choice theory, regarding the *equality* of treatment of alternatives and individuals:

- *Neutrality: For any profile $\boldsymbol{S}$ and permutation of entities $\rho : \mathcal{X} \to \mathcal{X}$, we have that $F(\boldsymbol{S}_\rho) = \rho(F(\boldsymbol{S}))$, where $\boldsymbol{S}_\rho$ is profile $\boldsymbol{S}$ with alternatives in $\mathcal{X}$ renamed by ρ.*

- *Anonymity: For any profile $\boldsymbol{S}$ and permutation of individuals $\rho : \mathcal{I} \to \mathcal{I}$, we have that $F(S_{\rho(1)}, \dots, S_{\rho(n)}) = F(S_1, \dots, S_n)$.*

Neutrality requires that if we rename the items in $\mathcal{X}$, the result should be the renaming of the initial result. Anonymity instead formalizes the fact that the collective opinion should not depend on the name of the individuals. The following two properties provide requirements on how to treat *consensus and total disagreement* in the individual preferences:

- *Weak-Pareto: If $\boldsymbol{S}$ is a profile in which all individuals rank a above b, then $b \notin F(\boldsymbol{S})$.*

- *Cancellation. If a profile $\boldsymbol{S}$ is symmetric then all entities are in the winning set, i.e., $F(\boldsymbol{S}) = \mathcal{X}$.*

The weak-Pareto property is a fundamental property when aggregating individuals' preferences: agreement among all individuals should be reflected in the collective opinion. Cancellation requires instead that if the disagreement is so extreme, as in a symmetric profile where individuals come in pairs whose preferences cancel each other out, then all items should be declared as the most preferred ones in the collective opinion.

The following two properties formalize the requirement that more information collected over an individual's preferences should lead to a result that is more preferred by that individual:

- *Voters participation: For all profiles $\boldsymbol{S} = (S_1, \ldots, S_n)$ and SP-structure S_{n+1}, any candidate in $F(\boldsymbol{S} + S_{n+1}) \setminus F(\boldsymbol{S})$ is ranked in higher or same position than any candidate in $F(\boldsymbol{S})$ in the preferences of voter $n + 1$.*

- *Rank participation: For all profiles $\boldsymbol{S} = (S_1, \ldots, S_n)$ and SP-structure $S' \subset S_n$, any candidate in $F(\boldsymbol{S}) \setminus F(\boldsymbol{S}_{-n} + S')$ is ranked in higher or same position than any candidate in $F(\boldsymbol{S}_{-n} + S')$ in the preferences of voter n.*[1]

In the classical voting theory context, participation means that voters have an incentive to participate (39). In a sentiment analysis context individuals have already expressed their opinion, so we do not need to favor their participation. However, the first property tells us that considering one more individual in the computation of the collective sentiment should result in a candidate that is higher in her ranking. In a similar way, the second property requires that more information on an individual's opinion lead to results that the individual ranks higher.

Finally, the following property formalizes the possibility of using techniques such as *map-reduce* (18) in particular profiles, for a more efficient computation of the set of most preferred alternatives:

- *Consistency: For all profiles $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$, if $F(\boldsymbol{S}_1) \cap F(\boldsymbol{S}_2) \neq \emptyset$ then $F(\boldsymbol{S}_1 + \boldsymbol{S}_2) = F(\boldsymbol{S}_1) \cap F(\boldsymbol{S}_2)$.*

Consistency can be an important property in an application domain where one needs to deal with big quantities of data, such as sentiment analysis. It tells us that if we manage to partition the (possibly very large set of) individual opinions into smaller sets which have some best candidate in common, perhaps by means of a proper heuristic, then we can work on the elements of the partition independently. Thus *divide and conquer* approaches are possible, which parallelize and possibly speed up the computation.

All properties presented above are adapted from the literature on social choice theory (4). Not all combinations of axiomatic properties are feasible: for instance, the well-known Arrow's Theorem showed that it is not possible to aggregate linear orders using a rule that satisfies three simple desirable properties (namely, a weaker version of anonymity, an additional property called independence of irrelevant alternatives, and weak-Pareto) (3). This is not the case for the list of axioms presented above, as we will show in the following sections.

4.2 The $B^*_{\underline{\alpha}}$ Rule

In this section we propose a parameterized class of aggregation procedures for profiles of SP-structures that builds on the classical Borda count, taking into account the

[1] Inclusions of SP-structures is defined as inclusions of preorders, and $\boldsymbol{S}_{-n}$ represents profile $\boldsymbol{S}$ without SP-structure S_n.

incompleteness of the ordering and the additional information given by the sentiment polarity expressed by the individuals.

Definition 10. *Given an SP-structure $S = (\mathcal{P}, \mathcal{N}, \mathcal{Z})$ over $\mathcal{X}$, the $s^*_{\underline{\alpha}}$*-score *of an entity $c \in \mathcal{X}$ in S is defined as follows:*

$$s^*_{\underline{\alpha}}(c, S) = \begin{cases} \alpha_1|\operatorname{down}^{\mathcal{P}}(c)| + \alpha_2|\operatorname{inc}^{\mathcal{P}}(c)| + \alpha_3|\mathcal{Z}| + \alpha_4 & \textit{if } c \in \mathcal{P} \\ -\alpha_1|\operatorname{up}^{\mathcal{N}}(c)| - \alpha_2|\operatorname{inc}^{\mathcal{N}}(c)| - \alpha_3|\mathcal{Z}| - \alpha_4 & \textit{if } c \in \mathcal{N} \\ 0 & \textit{if } c \notin \mathcal{P} \cup \mathcal{N} \end{cases}$$

Where $\underline{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ with $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}^+$. If c in $\mathcal{P}$, $\operatorname{down}^{\mathcal{P}}(c)$ *is the set of elements of $\mathcal{N}$ that are less preferred than c,* $\operatorname{up}^{\mathcal{P}}(c)$ *is the set of elements of $\mathcal{P}$ that are more preferred than c, and* $\operatorname{inc}^{\mathcal{P}}(c)$ *is defined as the the set of elements that are incomparable to c in $\mathcal{P}$ (in $\mathcal{N}$, respectively, for* $\operatorname{down}^{\mathcal{N}}$, $\operatorname{up}^{\mathcal{N}}$ *and* $\operatorname{inc}^{\mathcal{N}}$*). We will omit the reference to S when it is clear from the context.*

The $s^*_{\underline{\alpha}}$-score is defined as a parametrized class of scoring functions over SP-structures. It combines the approach from sentiment analysis, giving α_4 points to each alternative in the positive set and $-\alpha_4$ to all those in the negative set, with a generalization of the classical Borda rule, giving α_1 points to an alternative for all those that are ranked below, α_2 points for those ranked incomparable, and α_3 points for those alternatives that are in the neutral set (negative points if the alternative is in the negative set).

Note that no point is given to entities for which an individual has a neutral sentiment or for which she does not have any opinion. We are hence assuming that all score variables are initialized to 0, while an equivalent formulation could leave unspecified the score of alternatives for which no opinion has been extracted. The main difference between alternatives in $\mathcal{Z}$ and alternatives in $\mathcal{X} \setminus (\mathcal{P} \cup \mathcal{N} \cup \mathcal{Z})$ is that the former do contribute to the score of alternatives in the positive or in the negative set via the parameter α_3, while the latter are not taken into consideration when assigning scores to alternatives.

The use of a score may at first seems counterintuitive given our discussion in Section 3 on the interpersonal incomparability of preferential information. However, what is being represented in the $s^*_{\underline{\alpha}}$-score is purely ordinal information about the *number* of alternatives being more or less preferred to others, and should not be confused with the *intensity* of preference that could have been expressed in the individual data via the scoring function σ_i.

To exemplify the flexibility of our setting, we can consider several assumptions on the $\underline{\alpha}$ vector that the modeler can choose, depending on the application at hand. For instance, assuming $\alpha_1 > \alpha_2$ and $\alpha_3 > \alpha_2$ will make sure that more points are given to alternatives that are strictly preferred to others than to those that are incomparable. Another possibility is to assume that the score difference between two successive elements in the positive or negative part should be less than the score difference between the least positive and the best negative elements, for instance when $2\alpha_4 > \alpha_1$. If these two figures were equal, then the $s^*_{\underline{\alpha}}$-score would be equivalent to the classical Borda score when the set $\mathcal{Z}$ is empty, disregarding the sentiment information.

Definition 11. *The score of an entity $c \in \mathcal{X}$ in the profile of SP-structures $\boldsymbol{S} = \{S_i = (\mathcal{P}_i, \mathcal{N}_i, \mathcal{Z}_i) \mid i \in \mathcal{I}\}$ is defined as follows:*

$$S^*_{\underline{\alpha}}(c, \boldsymbol{S}) = \sum_{i \in \mathcal{I}} s^*_{\underline{\alpha}}(c, S_i)$$

where $s^*_{\underline{\alpha}}(c, S_i)$ is $s^*_{\underline{\alpha}}$-score of alternative c in the SP-structure S_i. The winners of the $B^*_{\underline{\alpha}}$ rule *are the candidates with maximal total score:*

$$B^*_{\underline{\alpha}}(\boldsymbol{S}) = \operatorname*{argmax}_{c \in \mathcal{X}} S^*_{\underline{\alpha}}(c, \boldsymbol{S})$$

Example 6. *Let the parameters in $\underline{\alpha}$ be $\alpha_1 = \alpha_4 = 2$ and $\alpha_2 = \alpha_3 = 1$, and therefore let the corresponding score be as follows*

$$s^*_{(2,1,1,2)}(c, S) = \begin{cases} 2 \times |\operatorname{down}^{\mathcal{P}}(c)| + |\operatorname{inc}^{\mathcal{P}}(c)| + |\mathcal{Z}| + 2 & \textit{if } c \in \mathcal{P}_i \\ -2 \times |\operatorname{up}^{\mathcal{N}}(c)| - |\operatorname{inc}^{\mathcal{N}}(c)| - |\mathcal{Z}| - 2 & \textit{if } c \in \mathcal{N}_i \\ 0 & \textit{if } c \notin \mathcal{P}_i \cup \mathcal{N}_i \end{cases}$$

The winner of $B^*_{(2,1,1,2)}$ on the profile of SP-structures associated with Example 2 is R. Indeed, the score $s^*_{(2,1,1,2)}(R) = 4$ since there are 2 elements ranked below R in the positive part by the first individual ($4 + 2$ points), and R is ranked in the negative side by individual 2 (-2 point). G follows with a score of 1 since there is one element incomparable in the positive part by the first individual (1 + 2 points), G is ranked in the positive side by individual 2 (+2 point) and G is ranked in the negative side by individual 3 with 2 elements in the neutral set (-2-2 points). B is the worst preferred alternative with a score of -1, obtaining 3 points by individual one, -4 points by the second individual and 0 points by the third individual.

4.3 Axiomatic analysis

We begin by showing that our Borda count generalizes the existing approaches used by sentiment analysis and preference aggregation. Let us first introduce some notation. Call a profile *purely preferential* if, for all $i \in \mathcal{I}$, the set $\mathcal{P}_i$ is equal to $\mathcal{X}$ and is linearly ordered, i.e., $\leqslant^{\mathcal{P}}_i$ is anti-symmetric, transitive and complete. Call a profile *purely sentimental* if for all $i \in \mathcal{I}$ the two sets $\mathcal{P}_i$ and $\mathcal{Z}_i$ form a partition of $\mathcal{X}$, and the candidates in $\mathcal{P}_i$ are all incomparable, i.e., the relation $\leqslant^{\mathcal{P}}_i$ is empty, and the set $\mathcal{N}_i$ is also empty. We now show that $B^*_{\underline{\alpha}}$ coincides with the Borda rule on purely preferential profiles, and that it coincides with approval voting on purely sentimental ones.

Theorem 1. *If a profile $\boldsymbol{S}$ is purely preferential, then for all $\underline{\alpha}$ we have that $B^*_{\underline{\alpha}}(\boldsymbol{S}) =$ Borda$(\boldsymbol{S})$. If a profile $\boldsymbol{S}$ is purely sentimental, then for all $\underline{\alpha}$ such that $\alpha_2 = \alpha_3$ we have that $B^*_{\underline{\alpha}}(\boldsymbol{S}) =$ Approval$(\boldsymbol{S})$.*

Proof. Let $\boldsymbol{S}$ be a purely preferential profile, i.e., all individuals are expressing a linear order over entities in $\mathcal{X}$ which are all in $\mathcal{P}$. Let $S^B(c, \boldsymbol{B})$ be the classical Borda score, i.e., the number of candidates ranked below c. Since in a purely preferential profile there are no alternatives that are incomparable to each other, and the sets $\mathcal{N}$ and $\mathcal{Z}$ are empty, the score $S^*_{\underline{\alpha}}(c) = \alpha_1 S^B(c) + \alpha_4 n$, where n is the number of voters. Since n is constant then the two rules elect the same candidates, no matter the value of α_1 and α_4.

Let now $\boldsymbol{S}$ be a purely sentimental profile and let $S^A(c)$ be the approval score of an entity c, i.e., the number of individuals approving c. Since all alternatives in $\mathcal{P}_i$ are incomparable, every approved entity in each single SP-structure gets a score equal to $\alpha_2(|\mathcal{X}| - 1) + \alpha_4$. To see this, it is sufficient to observe that alternatives in $\mathcal{Z}$ give α_3 points to alternatives in $\mathcal{P}$ and $\alpha_2 = \alpha_3$, and moreover in a purely sentimental profile the two sets $\mathcal{P}$ and $\mathcal{Z}$ form a partition of $\mathcal{X}$. Hence, we obtain that $S^*_{\underline{\alpha}}(c) = (\alpha_2(|\mathcal{X}|-1)+\alpha_4)\cdot S^A(c)$ and thus $B^*_{\underline{\alpha}}$ elects the same candidates as approval voting. □ □

Theorem 1 formalizes the fact that our proposed Borda count is a generalization of both approaches at the extreme of the spectrum described in Section 3: a pure sentiment analysis approach, which uses approval voting, and a pure preference aggregation approach, as described by the Borda rule. The result of Theorem 1 can be generalized to profiles of partial orders to show that $B^*_{\underline{\alpha}}$ extends the partial Borda count defined by Cullinan, Hsiao and Polett (16) as well as the *bucket averaging method* of Fagin et al. (24).

We now show that our proposed Borda count for collective sentiment analysis satisfies all the axiomatic properties presented in Section 4.1.

Theorem 2. *$B^*_{\underline{\alpha}}$ satisfies consistency, neutrality, anonymity, voters participation, rank participation, and cancellation for all $\underline{\alpha}$. If we assume that $\alpha_1 \geqslant \alpha_2$, then $B^*_{\underline{\alpha}}$ also satisfies weak-Pareto.*

Proof. For the sake of clarity we omit the reference to $\underline{\alpha}$ where it is not necessary. To prove that $B^*_{\underline{\alpha}}$ satisfies *consistency* it is sufficient to observe that $S^*_{\boldsymbol{S}_1+\boldsymbol{S}_2}(c) = S^*_{\boldsymbol{S}_1}(c) + S^*_{\boldsymbol{S}_2}(c)$. Those entities with maximal score in both $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are then the entities with maximal score in $\boldsymbol{S}_1 + \boldsymbol{S}_2$. *Neutrality* and *anonymity* are straightforward consequences of our definition of $S^*_{\underline{\alpha}}$.

A simple monotonicity argument can be used to prove both versions of *participation*. Consider first voters-participation. Let $w \in B^*_{\underline{\alpha}}(\boldsymbol{S})$, and let S_{n+1} be the additional SP-structure. We show that the winner of the joint profile w' is not worse in $n+1$'s ranking than w. Since we have only added information from $n+1$, w' must have received strictly more points than w to become the new winner, and this can only happen if agent $n+1$ prefers w' to w. The same argument can be straightforwardly adapted to the case of rank-participation.

Finally, to prove that $B^*_{\underline{\alpha}}$ satisfies *cancellation* we observe that in a symmetric profile all entities have score 0, since $S^*_{\underline{\alpha}}$ is symmetric with respect to $\mathcal{P}$ and $\mathcal{N}$.

For *weak-Pareto*, there are four cases for an individual to rank a above b, and we can show that in all cases $s^*_{\underline{\alpha}}(a) > s^*_{\underline{\alpha}}(b)$ and thus that b cannot be in the winning set. Recall that we assumed $\alpha_1 \geqslant \alpha_2$. Assume that $b \leqslant^{\mathcal{P}}_i a$ and $a \nleqslant^{\mathcal{P}}_i b$. If a third alternative c is ranked below b then by transitivity c is also ranked below a, and hence a and b get the same points from c. If $c \in \mathrm{inc}^{\mathcal{P}_i}(b)$ or $c \in \mathcal{Z}_i$ then this also gives the same points to a (or more, if c is ranked below a since $\alpha_1 \geqslant \alpha_2$). Finally, $\mathrm{down}^{\mathcal{P}_i}(b) \subset \mathrm{down}^{\mathcal{P}_i}(a)$ and thus a gets α_1 more points than b. The case in which a

and b are both in $\mathcal{N}_i$ is treated symmetrically: just consider alternatives ranked above the two and the set $\mathrm{up}^{\mathcal{N}_i}(a)$ rather than $\mathrm{down}^{\mathcal{P}_i}(a)$. If we instead assume that $a \in \mathcal{P}_i$ and $b \in \mathcal{Z}_i \cup \mathcal{N}_i$, then it is easy to observe that $s^*_{\underline{\alpha}}(a) > 0$ while $s^*_{\underline{\alpha}}(b) \leqslant 0$ and thus that also in this case $s^*_{\underline{\alpha}}(a) > s^*_{\underline{\alpha}}(b)$. Finally, if $a \in \mathcal{Z}_i$ and $b \in \mathcal{N}_i$ then $s^*_{\underline{\alpha}}(a) = 0$ but $s^*_{\underline{\alpha}}(b) < 0$ since $b \in \mathcal{N}_i$ gets $-\alpha_4$ points and any other alternatives $c \in \mathcal{N}_i$ can only decrease the score of b. □ □

We conclude this section by comparing $B^*_{\underline{\alpha}}$ with another rule, introduced in previous work by Brams and Sanver (9), that aims at combining approval voting with preference aggregation: *fallback voting*. Under this rule, each voter approves a subset (which could be empty) of candidates and ranks them in a linear order. The winner of fallback voting is obtained in an iterative way, by first checking whether there is a candidate that is top-ranked by a majority of voters. If such a candidate does not exist, then the first and the second ranked candidates in each voters' preference are considered, and once again it is checked if there is a candidate that is ranked first or second by a majority of the voters. The process goes on adding the third and subsequently ranked candidates until an alternative obtains a majority of approvals. The structures used by fallback voting to combine approvals with comparative preferences can be seen as a special case of SP-structures, with no neutral nor negative sets and linearly ordered items. Fallback voting may result in a different outcome than $B^*_{\underline{\alpha}}$. A detailed study of the difference between these two rules is left as future work. While fallback voting constitutes an interesting voting rule when individuals have incentives to express their preferences, in applications such as sentiment analysis the individual data that is collected will rarely be complete (see Section 3). Hence the need for rules that are able to handle incomplete profiles, such as our proposed $B^*_{\underline{\alpha}}$.

4.4 Algorithmic properties of $B^*_{\underline{\alpha}}$

In this section we analyze the algorithmic aspects of our proposed Borda count for collective sentiment analysis. Given the envisioned application, it is important that the basic problem of computing the most preferred alternative in a given profile be tractable, i.e. that the computational complexity of winner determination be solvable in polynomial time. We also provide an exact bound on the minimum number of bits required to compute the outcome of $B^*_{\underline{\alpha}}$ (aka. its communication complexity) and we show that it can be computed with an incremental algorithm.

Let us first make some considerations about the size of representing an SP-structure. Recall that we start from a set of m alternatives or entities $\mathcal{X}$, and n individuals. We assume that the sets $\mathcal{N}$, $\mathcal{P}$ and $\mathcal{Z}$ are encoded with a vector of length m containing for each alternative in $\mathcal{X}$ a label of 2 bits for the set to which it belongs to. Since the set $\mathcal{N}$ and $\mathcal{P}$ are disjoint, the two preorders $\leqslant^{\mathcal{P}}$ and $\leqslant^{\mathcal{N}}$ can be represented as binary relations on an $m \times m$ matrix, indicating for every pair (a, b) whether $a \leqslant^{\mathcal{P}} b$ or $a \leqslant^{\mathcal{N}} b$. The size of a profile of SP-structures with n individuals is therefore $O(nm^2)$.

The problem of *winner determination* is the algorithmic task of deciding whether a designed alternative $a \in \mathcal{X}$ is in the winning set of a given profile of SP-structures $\boldsymbol{S}$. This problem has been widely studied in voting theory (12), where its tractability is often considered a requirement for a rule to be considered of practical interest. We now show that the winner of our proposed Borda count can be computed in polynomial time:

Theorem 3. *The winner of $B^*_{\underline{\alpha}}$ can be computed in time $O(nm^2)$, hence in time linear in the size of the input.*

Proof. Given a single SP-structure S_i, the $s^*_{\underline{\alpha}}$-score of an alternative a can be computed in the following way. First check whether $a \in \mathcal{Z}_i$, $a \in \mathcal{P}_i$, or $a \in \mathcal{N}_i$, which can be done in constant time. If $a \in \mathcal{Z}_i$ then its score is 0. Otherwise we can compute its score by first counting how many alternatives are in $\mathcal{Z}_i$, which can be done in $O(m)$, and then counting how many alternatives are ranked below a, in case $a \in \mathcal{P}_i$, or how many alternatives are ranked above a, if $a \in \mathcal{N}_i$, and finally how many alternatives are incomparable to a. All these operations can be done in $O(m)$. We repeat this process for each alternative a and for each individual i, obtaining the upper bound $O(nm^2)$. □ □

A further important algorithmic property of a voting procedure is its *communication complexity*, i.e. the minimal amount of bits that needs to be expressed by the individuals in order to compute the most preferred alternatives. Previous work (15) provided lower and upper bounds for many voting procedures including the Borda rule, showing that the communication complexity of computing the Borda winner is $\Theta(nm \log m)$, i.e., lower and upper bounds are both equal, up to multiplication by a constant, to the function $nm \log m$.Since our Borda count for collective sentiment analysis generalizes the classical Borda rule (see Theorem 1), the communication complexity could in principle be higher, but as we show in the following result we simply move from a polylogarithmic to a polynomial factor:

Theorem 4. *The communication complexity of $B^*_{\underline{\alpha}}$ is in $\Theta(nm^2)$.*

Proof. An upper bound is easy to obtain, since it is sufficient for each individual to specify their SP-structure to be able to compute the winner of $B^*_{\underline{\alpha}}$. Given our representation of SP-structures, a profile can be specified in $O(nm^2)$ bits. A lower bound can instead be obtained by adapting the same bound for the classical Borda rule provided in (15). □ □

We conclude the section by proposing a notion of *incremental complexity* that should capture the feasibility of computing the result of an aggregation procedure in an on-line fashion. This is a very important aspect when data is examined incrementally or when using methods such as map-reduce (18) to deal with large quantities of data. Recall the two participation axioms we introduced in Section 4.1: they imply that the result of an aggregation procedure should take into consideration the additional information collected from the individuals. A good aggregation procedure that can be used to define a notion of collective sentiment should not only take care of this additional information, but also be able to update the outcome in little time.

Let $\mathcal{X}$ be a set of alternatives. An *individual expression* P over $\mathcal{X}$ is a preference, a vote, an opinion, or an SP-structure defined on $\mathcal{X}$. Given a profile of individual expressions $(P_1, \dots, P_n)$ – be it a profile of linear orders, of approval sets or of SP-structures – a *generalized voting rule* F is a function that outputs a set of most preferred alternatives $F(P_1, \dots, P_n) \subseteq \mathcal{X}$. If we denote with $\mathcal{PR}$ the space of all possible profiles for all finite n, then $F : \mathcal{PR} \to 2^{\mathcal{X}}$. The Borda rule, approval voting, fallback voting and $B^*_{\underline{\alpha}}$ are all generalized voting rules. We give the following definition:

Definition 12. *An generalized voting rule F is* incremental *if there exists a representation of profiles $r : \mathcal{PR} \to \{0,1\}^*$ and a function $\hat{F} : \{0,1\}^* \times \mathcal{PR}^1 \to 2^{\mathcal{X}}$, where $\mathcal{PR}^1$ is the set of all individual expressions, such that:*

- *$F(P_1, \dots, P_{n+1}) = \hat{F}(r(P_1, \dots, P_n), P_{n+1})$ for all profiles $(P_1, \dots, P_{n+1}) \in \mathcal{PR}$ and all finite n;*
- *for every sequence of individual expressions $\{P_i \mid i \in \mathbb{N}\}$, the following holds:*

$$\lim_{n \to +\infty} \frac{\textit{size}[r(P_1, \dots, P_n)]}{\textit{size}[(P_1, \dots, P_n)]} = 0$$

A generalized voting rule is incremental if its outcome, i.e. the set of most preferred candidates, can be computed by receiving the individuals expressions in a sequence, at each step computing the new outcome and storing a minimal amount of information that is needed to compute the outcome of the following step. Moreover, we require that the information stored at each step using function r be much smaller than the full representation of a profile as n grows.

Let us first show that both the Borda rule and approval voting are incremental. The Borda rule can be computed incrementally by storing the total Borda score of each alternative, and this can be done in space $O(m \log(nm))$ since $n \times (m-1)$ is the maximal Borda score that an alternative can obtain. When additional information is collected in the form of a linear order P_{n+1}, the total Borda scores can simply be updated with the additional scores computed. Since the size of a profile of linear orders is $O(nm \log m)$, the requirement on the size of the representation holds:

$$\lim_{n \to +\infty} \frac{\textit{size}[r(P_1, \dots, P_n)]}{\textit{size}[(P_1, \dots, P_n)]} = \lim_{n \to \infty} \frac{O(m \log(nm))}{O(nm \log m)} = \lim_{n \to \infty} \frac{\log n}{n} = 0$$

Approval voting receives as input a profile of sets of approved candidates, which has size $O(nm)$. An incremental procedure for its computation stores the number of approvals received by any candidate, and updates them when a new voter submits her ballots. Hence $\textit{size}[r(P_1, \dots, P_n)] = O(m \log n)$, showing that approval voting is also an incremental aggregation procedure.

Let us conclude by showing that $B^*_{\underline{\alpha}}$ is incremental for any $\underline{\alpha}$. In the same way as the Borda rule, the total score obtained by each alternative in a profile of SP-structures $\boldsymbol{S}$ can be stored in space $O(m \log(nm))$, since the maximal $S^*_{\underline{\alpha}}$-score is a linear function of $n \times m$. When a new SP-structure S is extracted, the total scores can be updated and the new outcome computed. Since a profile of SP-structures is represented in space $O(nm^2)$ we obtain the following:

$$\lim_{n \to +\infty} \frac{\textit{size}[r(P_1, \dots, P_n)]}{\textit{size}[(P_1, \dots, P_n)]} = \frac{m \log(nm)}{nm^2} = 0$$

We conjecture that all anonymous voting rules are incremental by Definition 12. In fact, if a voting rule is anonymous then the information contained in a profile with n voters can be summarized by using an amount of space that grows sub-linearly with n. However, our definition measures an important characteristic of a generalized voting rule, one that is particularly useful in applications that deal with large numbers of individuals. The notion of incrementality we proposed resembles the *on-line time* discussed by Maudet et Al. (36), which however focuses mostly on space complexity requirements. A complete study of the notion of incrementality in generalized voting rules is beyond the scope of this paper, but the above discussion may serve as a starting point for further work on this topic.

5 Empirical Analysis

This section reports on our experimental evaluation of the $B^*_{\underline{\alpha}}$ rule proposed in Section 4. The problem we face in this section is two-fold: First, in order to assess the relevance of preferential ordering information in determining the collectively preferred alternatives, we present experiments showing that there is a significant difference between the classical definition used by sentiment analysis and the result of the $B^*_{\underline{\alpha}}$ rule. Second, given the information sparsity which is characteristic of sentiment analysis domains, we evaluate the accuracy of the $B^*_{\underline{\alpha}}$ rule in situations of incompleteness, showing that its accuracy grows linearly with the amount of information available. In all our experiments we fix the parameters of the $B^*_{\underline{\alpha}}$ rule to $\underline{\alpha} = (2, 1, 1, 2)$.

5.1 Sentiment Analysis and Borda Count

A crucial factor supporting our claim that more complex models of preferences and aggregation procedures should be used in the definition of collective sentiment is that the classical sentiment analysis method (equivalent to approval voting) and $B^*_{\underline{\alpha}}$ output different results over the same data. In fact, if they did not differ enough, it would mean that the ordering information (not considered by approval/sentiment analysis) is not relevant for determining the winner. Thus there would be no point in extracting ordering information from individuals.

Figure 2 reports on our experiments on the simplest case of 2 candidates. We enumerated all profiles of totally ordered SP-structures with n voters, with n from 2 to 90, where a totally ordered SP-structure is an ordering over the two candidates (that is, a over b or b over a), plus a threshold which associates to each candidate either a positive or a negative sentiment. Thus, there are 6 possible such SP-structures. We have computed the winning candidates according to sentiment analysis (i.e, approval voting) and according to $B^*_{\underline{\alpha}}$, which in the case of 2 candidates is equivalent to using the majority rule, and we have counted the percentage of profiles on which the two winners are different. Figure 2 shows that such percentage stabilizes at around 30%.

We have also varied the number of candidates from 2 to 100, keeping the number of voters fixed at 10, in which case however we did not enumerate all possibilities but we generated 10.000 profiles of complete SP-structures with the impartial culture assumption, i.e., we sampled profiles with uniform distribution. Figure 3 shows that the

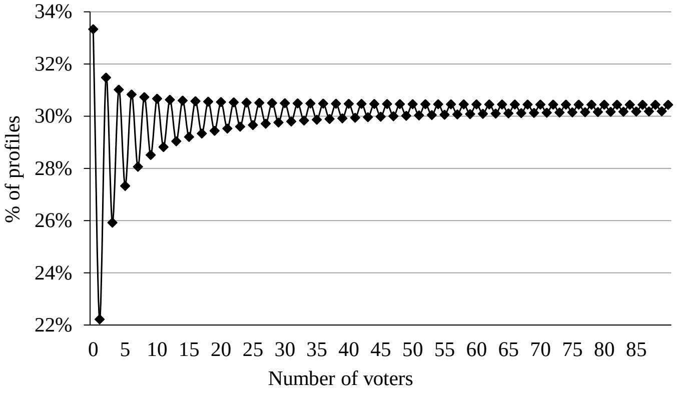

Figure 2: Percentage of profiles where sentiment analysis and $B^*_{\underline{\alpha}}$ differ (2 candidates).

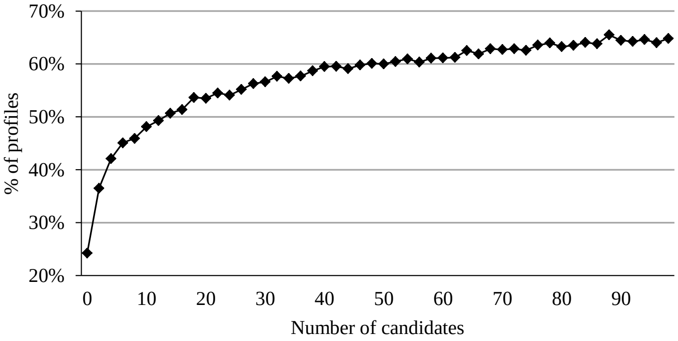

Figure 3: Percentage of profiles where sentiment analysis and $B^*_{\underline{\alpha}}$ differ (10 voters).

percentage of cases where $B^*_{\underline{\alpha}}$ yields a different result than sentiment analysis grows with the number of candidates reaching more than 60%.

5.2 Incomplete Data

In practical applications individuals are likely to express their opinions over a small subset of the alternatives under considerations, as observed, e.g., in the studies conducted on the Netflix dataset (7). It is therefore important to assess the behavior of our proposed Borda count on incomplete profiles.

To do this we generated profiles of complete SP-structures with 10 candidates and 100 voters, and we deleted a certain percentage of information to obtain an incomplete version of the profile. More precisely, we generated incomplete profiles in the following way: we first generated complete profiles and then we picked randomly a voter

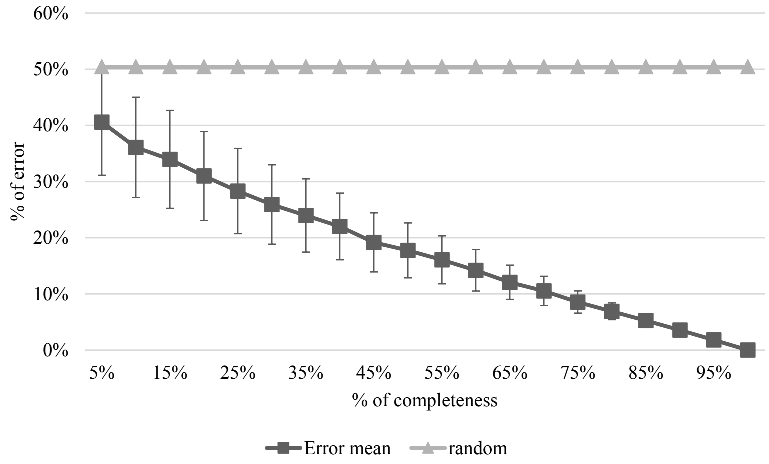

Figure 4: Mean error of $B^*_{\underline{\alpha}}$ on incomplete profiles in terms of $S^*_{\underline{\alpha}}$-score.

and a candidate, which is either positive or negative for that voter, and we changed the SP-structure of that voter to have no opinion on the selected candidate. With n voters and c candidates, nc corresponds to 100% of the information. Thus deleting $x\%$ of the information means performing the above described modification of the profile $(xnc)/100$ times. We then compared the winner (according to $B^*_{\underline{\alpha}}$) in the complete profile and in the incomplete one, by computing the absolute value of the difference between their $S^*_{\underline{\alpha}}$-scores, and we normalized it by dividing by the maximal error in the complete profile. Finally, we averaged over 10.000 profiles, obtaining the mean error introduced by the incompleteness of the profile. Figure 4 shows the trend in the error depending on the completeness of the profiles (mean error and variance). We also show the error of the random procedure, which outputs a candidate with uniform probability.

It is easy to see that Borda* always behaves better than the random procedure in identifying the winner in the complete profile, and moreover that its shape shows that accuracy quickly grows when the completeness of the profile increases.

6 Future Work

This work opens several directions for future work, and we conclude the paper by listing a number of challenges that arise from the use of techniques from preference aggregation and voting theory for collective sentiment analysis.

More refined models of opinions. As already noted in Section 2, our analysis of preference and opinion extraction disregarded two important parameters:

- *Time*. Individual opinions are expressed at a given point in time and are also subject to change or updates. Hence, temporal information plays an important role in defining a coherent individual view. We believe that the literature on knowledge representation (28), in particular belief revision and merging, provides useful tools for the analysis and summarization of conflicting information that can be applied to the modeling of this problem.

- *Features*. Entities or items are usually described by means of features, i.e., they may be elements of a product space. Techniques from natural language processing can be used to extract the relevant features and thus build the set of entities. However, in this setting preferences and opinions may compare features rather than entities, requiring a more elaborate framework for its extraction and representation. Moreover, the combinatorial explosion resulting from a large set of features may give rise to computational problems that require an adequate compact representation framework for preferences. The literature on social choice in combinatorial domains (31) and in particular on judgment aggregation (20) is highly relevant to this problem.

Validation of aggregation rules. Since the variety of preference aggregation methods that can be defined is very large, of which a prime example is the $B^*_{\underline{\alpha}}$ rule depending on the values given to its parameters $\underline{\alpha}$, a natural question is how to make a choice among them. Two options are possible, depending on the use of sentiment analysis techniques as a predictor for real-world events or as a tool for decision-making. First, if methods of collective sentiment analysis are used over time, tested for several settings and items, and employed in the context of predicting the result of real-world processes (such as elections or the evolution of a market, see, e.g., (2)), then *machine learning* techniques can be used to learn the best aggregation method, that is, the one that has proven to be the most accurate. Work on voting rules seen as *maximum likelihood estimators* can also be useful in this respect (15). Alternatively, as in classical voting theory and as performed in this paper, axiomatic properties as well as results about the computational complexity of aggregation rules could guide the choice of some aggregation methods over others.

Strategic behavior in sentiment analysis. The individuals composing a society, as well as the agents in a multiagent system, are very often connected by interpersonal ties, e.g., when individuals are organised in a network. In this case, individual preferences and opinions are not only the result of personal reflection but may also take into consideration the position taken by influential individuals or simply by agents that are close to them in the network. The field of *social network analysis* (29; 19) is a burgeoning research area which has the potential of generating highly interesting results once combined with techniques of preference and sentiment analysis. Sentiment analysis techniques are moreover not immune to strategic manipulation. A rising phenomenon is the creation of web services proposing the opening of thousands of fake Twitter accounts to be used as followers of the manipulator's account, or the publishing of big volumes of positive posts related to the manipulator's products. This represents a prime example of strategic behavior in collective choice problems, and the whole body of literature published on this topic may be put to test with real world data once the two fields of sentiment analysis and preference aggregation have been put together to their full potential. While the problem of manipulation for a single agent is computationally easy for the classical case of the Borda rule, we conjecture that for $B^*_{\underline{\alpha}}$ this is not the case, given the higher amount of possibilities that an agent has to manipulate the election. However, single-agent manipulation is unlikely to occur in sentiment analysis applications, given the high number of individuals concerned and the absence of a well-defined elicitation protocol. Instead, an interesting direction for future work is the study of coalitional manipulation, which was recently shown intractable even for the classical Borda rule (17; 8).

Big data and collective sentiment analysis. When the aggregation operation is relatively simple (e.g., the majority rule), it is possible to use straightforward techniques such as *Hadoop MapReduce* (18) to perform computations in parallel. The mapping phase can be used to run sentiment classifiers on text corpora in parallel; the resulting data objects can be combined/reduced in parallel. However, with more complex structures (e.g., conditional preference networks when the set of entities is described by means of features), the combination procedure may be more combinatorial in nature and may require non-trivial parallel processing. In this context, modern scale-out programming languages such as X10 can be particularly valuable (14), making it easy to write code that runs over thousands of cores and deals with hundreds of gigabytes of main memory data. Of particular interest is developing incremental parallel algorithms that can update collective sentiments as new utterances stream in and need to be processed.

References

[1] Ackerman, M., Choi, S.Y., Coughlin, P., Gottlieb, E., Wood, J.: Elections with partially ordered preferences. Public Choice **157**(1-2), 145–168 (2013)

[2] Arias, M., Arratia, A., Xuriguera, R.: Forecasting with twitter data. ACM TIST **5**(1), 8 (2013)

[3] Arrow, K.J.: Social Choice and Individual Values, second edn. John Wiley & Sons (1963)

[4] Arrow, K.J., Sen, A.K., Suzumura, K. (eds.): Handbook of Social Choice and Welfare. Elsevier (2002)

[5] Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of LREC-2010 (2010)

[6] Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls : Linking text sentiment to public opinion time series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (2010)

[7] Bennett, J., Lanning, S.: The Netflix prize. In: In KDD Cup and Workshop in conjunction with KDD (2007)

[8] Betzler, N., Niedermeier, R., Woeginger, G.J.: Unweighted coalitional manipulation under the borda rule is np-hard. In: Proceedings of IJCAI-2011 (2011)

[9] Brams, S., Sanver, M.R.: Voting systems that combine approval and preference. In: S. Brams, W. Gehrlein, F. Roberts (eds.) The Mathematics of Preference, Choice and Order. Springer (2009)

[10] Brams, S.J., Fishburn, P.C.: Voting procedures. In: K. Arrow, A. Sen, K. Suzumura (eds.) Handbook of Social Choice and Welfare. North-Holland (2002)

[11] Brams, S.J., Fishburn, P.C.: Approval voting, 2nd edn. Springer (2007)

[12] Brandt, F., Conitzer, V., Endriss, U.: Computational social choice. In: G. Weiss (ed.) Multiagent Systems. MIT Press (2013)

[13] Ceron, A., Curini, L., Iacus, S.M., Porro, G.: Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society (2013)

[14] Charles, P., Grothoff, C., Saraswat, V., Donawa, C., Kielstra, A., Ebcioglu, K., von Praun, C., Sarkar, V.: X10: an object-oriented approach to non-uniform cluster computing. In: Proceedings OOPSLA-2005 (2005)

[15] Conitzer, V., Sandholm, T.: Common voting rules as maximum likelihood estimators. In: Proceedings of UAI-2005 (2005)

[16] Cullinan, J., Hsiao, S., Polett, D.: A Borda count for partially ordered ballots. Social Choice and Welfare pp. 1–14 (2013)

[17] Davies, J., Katsirelos, G., Narodytska, N., Walsh, T.: Complexity of and algorithms for borda manipulation. In: Proceedings of AAAI-2011 (2011)

[18] Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)

[19] Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press (2010)

[20] Endriss, U.: Judgment aggregation. In: F. Brandt, V. Conitzer, U. Endriss, J. Lang, A. Procaccia (eds.) Handbook of Computational Social Choice. Cambridge University Press (2015)

[21] Endriss, U., Fernández, R.: Collective annotation of linguistic resources: Basic principles and a formal model. In: Proceedings of ACL-2013 (2013)

[22] Esuli, A., Sebastiani, F.: Sentiment quantification. IEEE Intelligent Systems **25**(4), 72–75 (2010)

[23] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E.: Comparing and aggregating rankings with ties. In: Proceedings of the 2004 ACM Symposium on Principles of Database Systems, pp. 47–58 (2003)

[24] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E.: Comparing partial rankings. SIAM Journal on Discrete Mathematics **20**, 47–58 (2006)

[25] Ganapathibhotla, M., Liu, B.: Mining opinions in comparative sentences. In: Proceedings of COLING-2008 (2008)

[26] Garg, A., Jayram, T.S., Vaithyanathan, S., Zhu, H.: Generalized opinion pooling. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (2004)

[27] Grandi, U., Loreggia, A., Rossi, F., Saraswat, V.: From sentiment analysis to preference aggregation. In: Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM-2014) (2014)

[28] van Harmelen, F., van Harmelen, F., Lifschitz, V., Porter, B.: Handbook of Knowledge Representation. Elsevier (2007)

[29] Jackson, M.O.: Social and Economic Networks. Princeton University Press (2010)

[30] Jindal, N., Liu, B.: Mining comparative sentences and relations. In: Proceedings of AAAI-2006 (2006)

[31] Lang, J., Xia, L.: Voting in combinatorial domains. In: F. Brandt, V. Conitzer, U. Endriss, J. Lang, A. Procaccia (eds.) Handbook of Computational Social Choice. Cambridge University Press (2015)

[32] Laslier, J.F., Sanver, M.R. (eds.): Handbook of Approval Voting. Springer (2010)

[33] Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer-Verlag New York, Inc. (2006)

[34] Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2012)

[35] Mao, A., Procaccia, A.D., Chen, Y.: Better human computation through principled voting. In: Proceedings of AAAI-2013 (2013)

[36] Maudet, N., Lang, J., Chevaleyre, Y., Ravilly-Abadie, G.: Compiling the votes of a subelectorate. In: Proceedings of IJCAI-2009 (2009)

[37] Metaxas, P., Mustafaraj, E., Gayo-Avello, D.: How (not) to predict elections. In: Proceedings of PASSAT-2011 and SOCIALCOM-2011 (2011)

[38] Mishne, G.: Predicting movie sales from blogger sentiment. In: In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW) (2006)

[39] Moulin, H.: Condorcet's principle implies the no show paradox. Journal of Economic Theory **45**(1), 53–64 (1988)

[40] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1-2), 1–135 (2008)

[41] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of EMNLP-2002, pp. 79–86 (2002)

[42] Pini, M.S., Rossi, F., Venable, K.B., Walsh, T.: Incompleteness and incomparability in preference aggregation: Complexity results. Artificial Intelligence **175**(7-8), 1272–1289 (2011)

[43] Rossi, F., Venable, K.B., Walsh, T.: A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers (2011)

[44] Xia, L., Conitzer, V.: A maximum likelihood approach towards aggregating partial orders. In: Proceedings of IJCAI-2011 (2011)

[45] Young, H.P.: An axiomatization of Borda's rule. Journal of Economic Theory **9**(1), 43–52 (1974)