MOX-Report No. 02/2017

Monitoring Rare Categories in Sentiment and Opinion Analysis - Expo Milano 2015 on Twitter Platform.

Arena, M.; Calissano, A.; Vantini, S.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Monitoring Rare Categories in Sentiment and Opinion Analysis - Expo Milano 2015 on Twitter Platform.

Marika Arena[a], Ana Calissano[b] and Simone Vantini[c]

January 13, 2017

[a] Dipartimento di Ingegneria Gestionale, Politecnico di Milano
Via Lambruschini, 4/B, I-20156 Milano, Italy
E-mail: marika.arena@polimi.it
[b] MOX, Dipartimento di Matematica, Politecnico di Milano
Piazza Leonardo da Vinci 32, I-20133 Milano, Italy
E-mail: anna.calissano@polimi.it
[c] MOX, Dipartimento di Matematica, Politecnico di Milano
Piazza Leonardo da Vinci 32, I-20133 Milano, Italy
E-mail: simone.vantini@polimi.it

Abstract

This paper proposes a new aggregated classification scheme aimed to support the implementation of text analysis methods in contexts characterised by the presence of rare text categories. The proposed approach starts from the aggregate supervised text classifier developed by Hopkins and King and moves forward relying on rare event sampling methods. In details, it enables the analyst to enlarge the number of text categories whose proportions can be estimated preserving the estimation accuracy of standard aggregate supervised algorithms and reducing the working time w.r.t. to unconditionally increase the size of the random training set. The approach is applied to study the daily evolution of the web reputation of Expo Milano 2015, before, during and after the event. The data set is constituted by about 900,000 tweets in Italian and 260,000 tweets in English, posted about the event between March 2015 and December 2015. The analysis provides an interesting portray of the evolution of Expo stakeholders' opinions over time and allow to identify the main drivers of Expo reputation. The algorithm will be implemented as a running option of the next release of R package ReadMe.

1 Introduction

From the 1st of May 2015 to the 31st of October 2015, Milano hosted the 2015 World Exposition (Expo Milano 2015). The event was initially characterized by doubts and uncertainties. The enthusiasm of hosting a world fair was accompanied by controversies concerning its organization; the opportunity of exploiting positive externalises induced by the event were strictly intertwined with the long-lasting discussion about the investments required to face its preparation. Newspapers ran reports of corruption episodes, cost overruns and delays. However, when the exposition started, initial skepticism gave way to growing curiosity and, in the end, turned out in an unexpected success. Milano Expo 2015 involved 140 countries and was visited by 21 millions of people, with 7 millions of foreign visitors and 2 millions of students (Expo S.p.a., 2015). The theme of the exposition, "Feeding the Planet, Energy for Life" marks an opportunity to put the centrality of sustainability at the top of the political agenda and stimulated visitors with thought-provoking ideas coming from the pavilions of different countries. But how did the reputation of Expo Milano 2015 evolve before, during, and after the event? Which are the main drivers influencing these changing dynamics?

To answer to these questions, we study the web reputation of Expo Milano 2015, by analysing Twitter data through sentiment and opinion analysis. The bustle that surrounded the Expo was mirrored in on-line discussions and web participation, making social media an interesting channel for understanding what people was thinking and saying about the Expo. Among the existing social media platforms, we focus on Twitter for two main reasons. Firstly, it has a public philosophy and via API, Twitter offers a partial free download of its data. Secondly, it is micro-blogging platform, where the users share in 140 characters their own opinion about specific topics. The sharpness of posts helps the sentiment analysis performances, conducted on sentence-level data-set.

We construct the data set downloading the tweets with tags related to Expo Milano 2015 covering the time-frame from the 17th of February 2015 to the 31st of December 2015. Because of the international vocation of the event, both Italian and English written Tweets are analysed. Figure 1 shows the amounts of analyzed Tweets (here aggregated per month), in Italian and English, respectively.

However, to fully capture the evolution of sentiment about Expo, we have to deal with a critical methodological issue, i.e.: the management of rare categories in the data set. As the mission of an Expo is educating

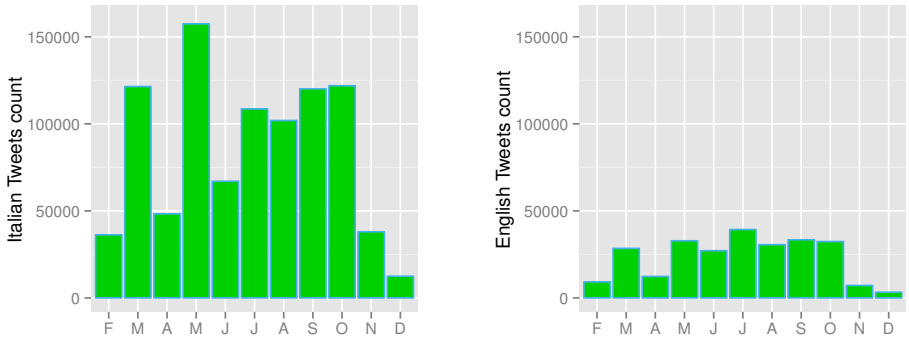


Figure 1: Downloaded Tweets from the 17th February 2015 to 31st December 2015 via keywords concerning Expo Milano 2015. Data are here represented monthly aggregated, in both Italian and English languages.

the public, sharing innovation, promoting progress and fostering cooperation among participating countries, the event put together many different stakeholders, moved by diversified expectations and perceptions, resulting in a complex and varying arrangement of interests and feelings. This heterogeneity was reflected in the on-line discourse, that was characterised by some "mainstream" topics discussed by plenty of people and some "less represented" categories - hereafter named rare categories - related to issues discussed by fewer people, but still relevant to understand the multifaceted reputation of the event.

The presence of rare categories is particularly critical for the implementation of supervised sentiment classifiers, that however, represent an essential instrument for performing sentiment analysis. As discussed in details in Section 3, supervised sentiment classifiers require a training set. The language used in the training set is assumed to be representative of the entire text (e.g., Hand, 2006), and it is labelled through hand-coding to obtain a better interpretation of the sentiment (e.g., Hopkins and King, 2010). When a corpus of texts is characterised by the presence of rare categories, there is a non-null probability of not gathering any text belonging to these rare categories in the training set, with the risk of losing some relevant pieces of information. Against this background, in this work we propose a new aggregated supervised classification scheme for sentiment and opinion analysis that takes advantage of the integration of standard sentiment and opinion analysis techniques with rare event sampling techniques. This approach allows the estimation of both broad-discussed and niche topics, contrary to

current approaches which are able to deal with the former ones exclusively. This specific feature is particularly relevant from a managerial point of view because the identification and the analysis of rare categories could be used to anticipate future trends, and to identify and manage potential risks or opportunities.

The rest of the paper is organized as follows. Section 2 outlines the state of the art about opinion mining, with particular attention to classification methods and, more specifically, aggregate supervised ones, that represent the starting point for this work. Section 3 introduces the proposed classification scheme and details it in terms of sentiment categories definition, texts pre-processing, variables definition, classification scheme evaluation, and results computation. Section 4 is fully dedicated to the analysis of the web reputation of Expo Milano 2015. Section 5 reports the results of a statistical comparison performed between our classification scheme and other existing ones. Finally, some points of discussion are reported in Section 6.

2 State of The Art

The concept of *sentiment analysis*, as described in Das and Chen (2001), indicates automatic analysis of evaluative texts and measures predictive judgement in it. Meanwhile, the concept of *opinion mining* was firstly introduced by Dave, Lawrence, and Pennock (2003), referring to processing a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them. Broadly speaking, sentiment analysis and opinion mining denote the same field of study and they are used interchangeability (Pang and Lee, 2008). Beside these general definitions, different methodologies have been developed, relying on similar input data (i.e., a corpus of texts) and sharing the same goal of enabling opinion-oriented information-seeking systems (Pang and Lee, 2008; Ribeiro et al., 2016). The conceptual taxonomy of the methods described hereafter follows the one introduced by Grimmer and Brandon (2013). The structure of this taxonomy is summarized in the following tree and detailed in rest of the section:

- Scoring methods
- Classification methods
  - Classification methods with unknown categories
  - Classification methods with known categories

- Dictionary methods
- Supervised methods
  - Individual
  - Aggregated (our proposal)

At a first level of analysis, these methods can be distinguished into: *scoring* and *classification* (Grimmer and Brandon, 2013). Scoring methods locate texts in an ideological space, weighting and sorting the words belonging to the document (e.g., Monroe and Maeda, 2004; Slapin and Proksch, 2008; Laver, Benoit, and Garry, 2003; Martin and Vanberg, 2008; Lowe, 2008). Classification methods, instead, organize texts into a set of categories, supposing that the meaning (and sentiment) of the text is given by considering the word combination (Grimmer and Brandon, 2013). In our study, we focus on classification methods since the segmentation of the corpus of texts into categories is more likely to provide a meaningful managerial interpretation of web reputation dynamics.

In classification methods, categories can be either *unknown* or *known* beforehand. Examples of methods relying on unknown categories are provided by Latent Dirichlet Allocation - LDA, which is a Bayesian generative model that encodes problem-specific structure into an estimation of categories (Blei, 2012), and several evolutions of this basic model (e.g., Blei et al., 2003; Blei, and Lafferty, 2012; Rao et al., 2014; Salter-Townshend and Murphy, 2014; Roberts, Brandon, and Airoldi, 2016). On the other hand, the goal of methods with categories known a priori is to assign data into these categories. The idea of the method can be based either on case specific dictionary, linking words to categories, or on supervised machine learning, or on new method which mix dictionary and learning (e.g., Taboada et al., 2011; Zhou, Zhang, and Sanderson, 2014; Mudinas, Zhang, and Levene, 2012). In this study, we selected the second typology of methods, because most of the factors that can potentially drive the web reputation of an event are known upfront, based on managerial literature.

Then, two different approaches can be further distinguished: *dictionary* methods, also called lexicon-based method, and *supervised* methods, also called learning-based methods. Dictionary methods are most intuitive and easy to apply for content analysis, because they assign texts to classes by using the frequency of established keywords (Stone et al., 1968; Rao et al., 2014; Nirmala, Roopa, and Naveen Kumar, 2015; Zhao et al., 2016; Choi and Pankoo, 2013). In supervised methods instead, the algorithm assigns texts into categories based on a predictive function, developed and learnt based on a training set, i.e., a subset of texts previously assigned within the

categories. Training set labels can be assigned either manually, by a human coder who reads and names the correct sentiment category, or automatically, using dictionaries, or other parameters into the texts such as positive and negative emoticons (Go, Bhayani, and Huang, 2009). The choice of tagging method is strictly problem-driven. Due to the complexity of our application, the large number of topics discussed on Twitter, and their heterogeneity we pursued an approach based on a supervised classification scheme trained on a manually hand-coded set.

Supervised classifiers can be *individual* or *aggregated*. Individual approach estimate the category of each texts in a new corpus of texts aiming at minimizing the probability of error in individual class assignment, while aggregated approach estimate the class proportions into the new corpus of texts aiming at minimizing the error between estimated proportions and true proportions. Some of the major individual supervised classifiers are listed and briefly described in Pang, Lee, and Vaithyanathan (2002) and Mukherjee and Pushpak (2013). For example, Random Forest (Breimen et al., 1984) is based on decision trees, Naive Bayes (Duda and Hart, 1973) is using Bayesian rule to assign the sentiment category which the text belongs to, Maximum Entropy (Berger, Della Pietra, and Della Pietra, 1996) maximises the entropy of a distribution with the parameters suggested by the training set distribution. Finally, Support Vector Machines (Joachims, 1998) are probably the most used method in this stream of literature. It is based on the optimization of the hyperplane separating document vectors belonging to different sentiment groups. Individual supervised classifiers have been generating soaring attention under different aspects, especially in terms of performance comparisons or slight variations and step forward or mixture of these well-known classifiers (da Silva, Hruschka, and Hruschka Jr, 2014; Mahalakshmi and Sivasankar, 2015; Tripathy, Agrawal, and Rath, 2016; Tian, et al., 2016; Erosheva, Fienberg, and Lafferty, 2004). Aggregate classifiers are more recent and have been introduced in Hopkins and King (2010). Their idea is to overcome the *classify-and-count* paradigm by directly estimating the aggregated proportions of texts associated to a certain sentiment, without assembling the results of several individual classifications. In this way, the method reduces both the number of steps to be performed for the computation of the sentiment and the mismatch between estimated proportions and true proportions (Ceron, Curini, and Iacus, 2013; Corallo et al., 2015). Another example of aggregate classifier developed from this seminal paper is the iSA method (Ceron, Curini, and Iacus, 2016). In our study, we clearly rely on an aggregated perspective since the focus is on accurately describing the overall perceived web reputation of the event with

little scope on describing the single user's opinion.

When a supervised aggregated classifier is used to address a complex problem, that is characterised by different topics, with frequent and less frequent occurrences, a complete and detailed training set is required, in order to cover their diversity and nuances. However, the classic training set obtained by random sampling (Hand, 2006) could hardly include examples of all the rare categories belonging to the data set, thus giving to the classifier non chance of "learning" the typical wording of texts expressing those opinions. To address this problem, scholars suggest to unconditionally increase the size of the random sampled training set (e.g., Hopkins and King, 2010). Still, this could result in higher time cost – due to the manual tagging – without ensuring the coverage of all the rare categories, because increasing the size of the random sample can only slightly increase the probability of collecting some texts belonging to rare categories. Moving from this consideration, we propose a new aggregated supervised classification scheme (namely, *Rare-but-not-least*), which is giving the chance of estimating all the interesting sentiment categories, including the rare ones. As detailed in the next session, Rare-but-not-least can be trained on a training set artificially reinforced with texts of specific categories (i.e., rare categories) still providing unbiased estimates of all category proportions. This fact simultaneously enables a detailed estimation of all categories of interest and reduces the time needed for hand-coding (i.e., the training set can be enlarged focusing on texts belonging to rare categories exclusively).

3 Methods: the Rare-but-not-least classification scheme for sentiment and opinion analysis

As outlined in Section 2, our method starts developing from the work of Hopkins and King (2010), which is considered the golden standard in sentiment analysis. From an application point of view, the innovative part of the approach lies into the possibility of moving from random sampling strategy of the training set, to biased sampling strategy, driven by case-specific categories. Random sampling does not indeed ensure the presence of texts belonging to rare categories, and causes an inefficient (and possibly useless) use of time in tagging redundant texts. To solve this problem, a natural possibility is of course to enrich the poll of texts belonging to rare categories. This type of sampling technique is strictly linked with the strategies known as *choice-based*, and *case-control* sampling (Breslow, 1996). Those strategies suggest to randomly select endogenous variables within categories of

exogenous variable, when one of the values of the exogenous variable is rare in population. In particular, we focus on the sampling solution proposed by King and Zeng (2001). Reported on our classification scheme, categories are interpreted as the exogenous variable and the texts as the endogenous variable (Hopkins and King, 2010), so texts are sampled conditioned to the categories they belongs to, via keywords searching. Nevertheless, this keywords driven sampling violates a crucial assumption which the approach of Hopkins and King (2010) is based on, i.e., the fact that the "words distribution" in each category of the training set should reflect the "word distribution" in the target population. For this reason, a naive training of the Hopkins and King's approach with a keywords reinforced training sample would lead to biased estimates of category proportions in the population of texts.

Thus, starting from the same model structure, our approach consists in a new unbiased classification approach which - differently from any known classifier - is trained by two training sets: (*i*) a standard unbiased corpus of texts obtained by random sampling (*ii*) and a biased one which is built ad-hoc by enriching through a keywords-based search the categories which present null or very small sample size in the former one. In the following, we will present the Rare-but-not-least classification scheme, detailing the traditional steps performed, when a supervised aggregate classification of texts is in order:

Step 1: Category definition, training set creation, and text stemming;

Step 2: Building the classifier;

Step 3: Performance evaluation;

Step 4: Estimation of category proportions on the target population.

3.1 Step 1: Category definition, training set creation, and text stemming

Step 1 of the procedure is detailed in the following and schematically summarized in Figure 2.

3.1.1 Category definition

In text analysis, categories can be conceptually divided into layers: sentiment and opinion ones. The sentiment expressed into the text is the speaker aptitude toward the issue object of investigation. Typical sentiments are

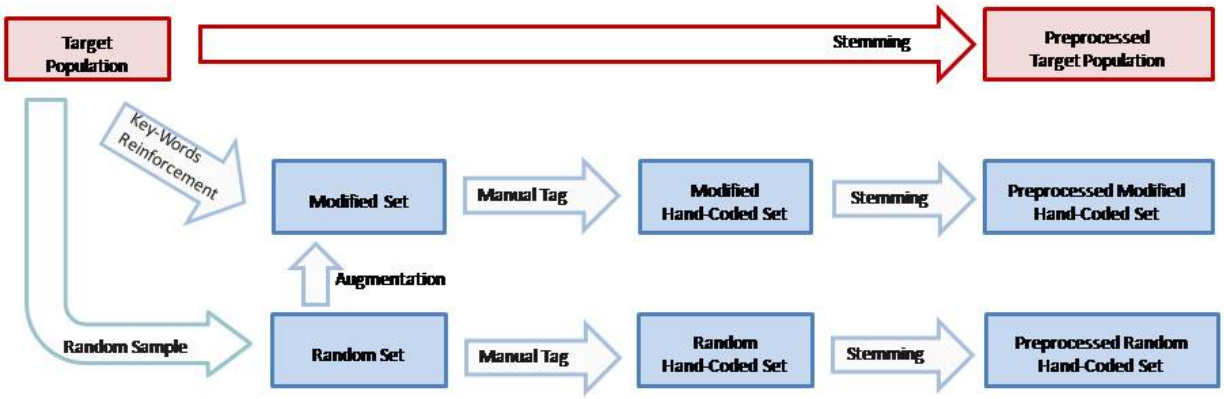


Figure 2: Step 1. Training set creation and stemming process. From the sampling strategy, through the manual tagging, to the ready to use matrices. Red boxes refer to the target population, while blue ones to the training sets.

"Positive", "Negative", "Neutral" and "Off topics". Opinion analysis goes instead deeper into the text, trying to detail also the reason why a text sentiment is positive, negative, or neutral. In opinion analysis categories are always more then the sentiment analysis ones and very case-specific.For both sentiment and opinion analysis, the definition of the categories is a crucial step since they will directly impact on the output accuracy. From this perspective, they need to be (e.g., Hopkins and King, 2010):

- mutually exclusive: each text cannot be associated to two or more categories;
- exhaustive: each text is associated to one category;
- relatively homogeneous: each category is ideally associated to a unique sentiment or opinion.

From a practical point of view, the common approach is to define all the possible interesting sentiment or opinion categories and possibly merge few of them if the properties above are not properly satisfied.

3.1.2 Training set creation

Creating a training set consists in collecting a sub sample of texts and assigning (typically manually) to each of them two labels, indicating respectively the sentiment and the opinion category. In retrospective studies (like the

one object of this manuscript), the *Random Hand- Coded set* (i.e. the classical training set) is obtained by randomly sampling texts along the entire duration of the study, such to capture all the possible language changes in the investigated time-frame. Not rarely, random sampling might lead to the absence (or near absence) of texts belonging to rare but important categories. A typical scenario where this issue is massively impacting is in the analysis of negative opinion categories in an event reputation study. To face this shortage, scholars suggest increasing the sample size by tagging new randomly sampled texts until a desired number of texts belonging to the rare categories is achieved (Hand, 2006). Differently, we hereby propose a new keyword-based sampling to obtain a second reinforced training set (i.e, the *Modified Hand-Coded set*) in which rare categories are efficiently reinforced. The Modified Hand-Coded set is shaped by adding to the Random Hand-Coded texts presenting keywords which are expected to appear with large probability in texts belonging to the rare categories. As a consequence, we have two major effects: (i) the distribution of rare categories in Modified Hand-Coded Set is of course over-estimated with respect to the Random Hand-Coded Set (this is not an issue for standard methods as we will show later); (ii) the distribution of words within each category will be biased toward an over-use of the selected keywords and other keyword-related words, this instead violates the assumption which standard methods are built on. We indeed hereby propose a method to correct this bias.

3.1.3 Text Stemming

In sentiment and opinion analysis, there are two groups of random variables which the classification scheme is dealing with: a former group made of a unique multinomial variable D modelling the text category and a latter set S of variables modelling the text wording. The first variable D trivially represents the random sentiment or opinion category of the text. In detail, D can assume the values $D_1, \ldots, D_J$ indicating the possible J text categories. Note that for all texts in the corpus to be classified (i.e., the target population) the values of the variable D is not directly observable.

Moving to the second group of variables, in order to statistically analyse texts, each text is preprocessed via standard procedures, until it becomes a binary vector. All texts are converted into lowercase letters and punctuation, extra white spaces, word prefixes and suffixes are removed. Thus, every word belonging to the same root is converted into the common *stem* (e.g. “read”, “reader”, “reading”, and “readable” are all changed into “read”). Finally, all the preprocessed texts are translated into a sequence binary

variables associated to the presence/absence in the text (i.e., 1/0) either of a stem (called "unigram"), two linked stems ("bigram") or even more than two stems linked together. In the current work, we will just rely on unigrams. Thus, the Random (or Modified) Hand-Coded and also the corpus of texts to be classified (i.e., the target population) are now converted into matrices of ones and zeros with as many rows as the number of texts and as many columns as the number K of stems used in the analysis. Note that typically both too frequent and too little frequent stems are not taken into consideration because of their little information content. Therefore, the number of columns K is typically much smaller than the actual amount of stems appearing in the analyzed texts. The ith row of each matrix is thus modelled as an instance of the random binary vector S which can assume $S_1, ..., S_{2^K}$ possible values (i.e., *word-stem profiles*) and whose distribution is dependent on the category which the text belongs to (i.e., the corresponding realization of the random variable D).

3.2 Step 2: Building the Classifier

Aggregated classification schemes are based on a direct relationship between category probabilities (our final goal) and word-stem profile probabilities, bypassing any text-specific category prediction. Specifically, let us consider the column vector $P(S)$ indicating the probabilities that each one of the 2^K possible word-stem profiles is observed in a text belonging to the target population and the column vector $P(D)$ indicating the probabilities that the same text belongs to the J possible categories. The two probability vectors are trivially linked through the following system of probabilistic identities:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P(S|D)} \underset{J \times 1}{P(D)}, \tag{1}$$

with $P(S|D)$ being a matrix, where each column represents the conditional probabilities of the 2^K possible word-stem profiles given the text categories, which can be interpreted as the probability that a certain combination of stems is used in the text given the sentiment or opinion of the text.

Hopkins and King proposal is to compute $P(S)$ directly on the target population (i.e., the corpus of texts to be classified), estimate $P(S|D)$ from a small sample of hand-coded texts, and finally estimate $P(D)$, consistently, relying on equations (1). Their proposal is based on two pragmatic considerations: (i) while $P(S)$ can be directly and automatically estimated by scrolling the corpus of texts and computing the relative frequency of each word-stem profile in the target population, on the contrary the direct estimation of $P(D)$ (which is our final goal) would require a full manual labelling

of the texts in the target population, which is instead not feasible; (*ii*) while $P(D)$ (and consequently $P(S)$) are clear dependent on the corpus of texts under investigation (for instance they naturally change even abruptly along time or across different information contexts), on the contrary changes in the matrix $P(S|D)$ are much slower along time and more stable across different information contexts. They basically assume that the velocity which people change their sentiment or opinion is extremely faster than the velocity which people change the way they express the same sentiment or opinion, i.e.:

$$P^{RH}(S|D) = P(S|D) \ , \tag{2}$$

where $P^{RH}(S|D)$ and the $P(S|D)$ indicate the conditional distribution computed on Random Hand-Coded set and on the target population, respectively. Since random sampling from the target population is a sufficient condition for (2), by replacing eq. (2) in eq. (1) we obtain:

$$\underset{2^K \times 1}{P(S)} = \underset{2^K \times J}{P^{RH}(S|D)} \underset{J \times 1}{P(D)} \ . \tag{3}$$

In detail, $P(S)$ is obtained by computing the relative frequencies of the word-stem profiles in the target population, $P^{RH}(S|D)$ by computing the relative frequencies of the word-stem profiles in each sentiment or opinion category in the Random Hand-Coded set, and finally $P(D)$ is consistently estimated by solving the $2^K \times J$ linear system (3) where typically $2^K >> J$. The generalized solution is simply obtained by minimizing the sum of the squared residuals of the equations, thus leading to:

$$P(D) = \left[P^{RH}(S|D)' P^{RH}(S|D)\right]^{-1} P^{RH}(S|D)' P(S) \ . \tag{4}$$

Our proposal is based instead on a new estimator of $P(S|D)$ gathering information from both the Random Hand-coded set and the Modified Hand-Coded set. The goal is still to obtain unbiased estimates of the conditional probabilities of all categories present in the Random Hand-coded set but also unbiased estimates of the conditional probabilities of possibly important rare sentiment or opinion categories present in Modified Hand-Coded set but not in the Random Hand-coded set. For this reason we named our classification scheme *Rare but not least*. Note that, as explained in Subsection 3.1.2, a naive replacement of the Random Hand-coded set with the Modified Hand-Coded set would violate assumption (2) thus providing biased estimates of $P(D)$.

In detail, our approach is based on the definition of two sets of correcting coefficients able to remove the bias in the estimates $P^{MODH}(S|D)$

obtained by directly computing the relative frequencies of the word-stem profiles in each sentiment or opinion category in the Modified Hand-Coded set. This bias correction is derived from the comparison of the Modified Hand-coded set with the Random Hand-Coded set. In detail, 2^K coefficients to rescale the rows and J coefficients to rescale the columns of the matrix $P^{MODH}(S|D)$ are computed.

Our proposal is based in detail on the linguistic assumption:

$$P^{RH}(D|S) = P^{MODH}(D|S) \tag{5}$$

which basically assumes that given a text (specifically, a word-stem profile), the probability that the text is expressing a certain sentiment or opinion does not depend on the text sampling strategy used to select the text. Note that, on the contrary, keyword-based sampling is likely to affect the word-stem profiles distribution within categories for the two sampling strategies, (i.e. $P^{RH}(S|D) \neq P^{MODH}(S|D)$) thus leading to the violation of the assumption (2). In detail, for every possible word-stem profile S_i with $i = 1, \ldots, 2^K$, the following coefficient is computed:

$$A_i = \frac{P^{RH}(S = S_i)}{P^{MODH}(S = S_i)} \ , \tag{6}$$

with $P^{RH}(S = S_i)$ and $P^{MODH}(S = S_i)$ indicating the relative frequency of the ith word-stem profile in the Random Hand-Coded set and in the Modified Hand-Coded set, respectively. Then, for every category $j = 1, \ldots, J$, the following coefficient is computed:

$$B_j = [\sum_{i=1}^{2^K} P^{MODH}(S = S_i|D = D_j)A_i]^{-1} \ , \tag{7}$$

with $P^{MODH}(S = S_i|D = D_j)$ indicating the relative frequency in the Modified Hand-Coded set of the ith word-stem profile in the texts belonging to the jth category. Finally, the following theorem basically provides an alternative way to estimate $P(S|D)$ under the assumptions (2) and (5). Its proof is reported in Appendix A.

Theorem 1. *Let A be a squared $(2^K \times 2^K)$ diagonal matrix with coefficients A_i on its diagonal and B a squared $(J \times J)$ diagonal matrix with coefficients B_j on its diagonal. If $P^{RH}(S|D) = P(S|D)$ and $P^{RH}(D|S) = P^{MODH}(D|S)$ then:*

$$AP^{MODH}(S|D)B = P(S|D) \ . \tag{8}$$

Consistently with Theorem (1) we hereby estimate the sentiment or opinion proportions in the target population as:

$$P(D) = \left[\left(AP^{MODH}(S|D)B \right)' \left(AP^{MODH}(S|D)B \right) \right]^{-1} \left(AP^{MODH}(S|D)B \right)' P(S) \,. \tag{9}$$

The great advantage of this alternative estimation is to leave space for a keyword-based reinforcement of the original random sample, which is a simple and efficient way to intercept texts belonging to specific (e.g. rare but important) text categories thus significantly enlarging the number of categories whose proportions can be estimated in the target population. Typical examples are categories that can negatively affect reputation. Texts of those categories are indeed rarely observed under normal conditions and thus weakly present (or totally absent) in randomly sampled sets. Nevertheless a monitoring of these categories is paramount for the early detection of possible spurts of these sentiments or opinions.

From a theoretical point of view, it is easy to see that when the modified sample coincide with the original sample both A and B are identity matrices. Thus, in this limit case, the estimates obtained by means of (9) coincide with the estimates obtained by means of the Hopkins and King method.

3.3 Step 3: Performance Evaluation

Once the categories are defines, the original and the modified training sets hand-coded, and the classification scheme implemented, a performance evaluation is run before estimating the category proportions in the target populations. A standard way to measure performance is dividing the Hand-Coded set in two parts, a large one acting as a labelled *training set* and a smaller one playing the role of the unlabeled *test set*. To measure the mismatch between the actual proportions of categories in the test set and the estimated proportions, we rely on the index commonly used by scholars (e.g. Hopkins and King (2010)) which basically computes the average squared difference between the actual proportions of categories in the test set and the estimated proportions over the J categories.

$$I = \sqrt{\frac{\sum_{j=1}^{J} \left[P_{true}(D_j) - P_{est}(D_j) \right]^2}{J}} \tag{10}$$

This index can be either used to compare different estimation methods for a given set of categories, or also to perform category selection (e.g., aggregation, splitting, addition, or removal of the original categories) for a given estimation method.

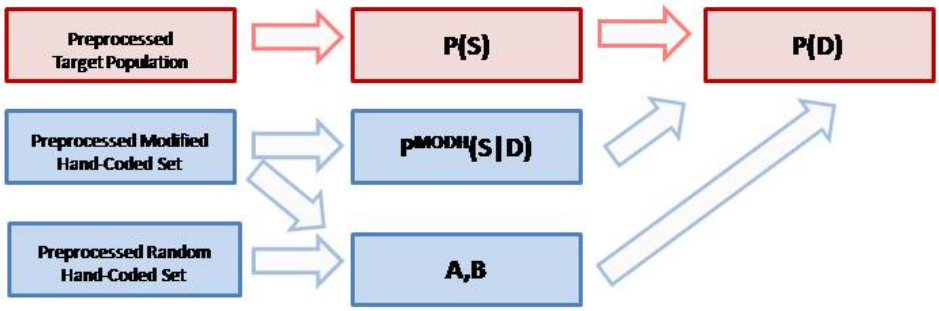


Figure 3: Estimation strategy scheme: from pre-processed texts to the final result. Red boxes refer to the target population quantities, while blue ones to the training sets ones.

3.4 Step 4: Estimation of category proportions on the target population

At this point to estimate the proportions of the selected categories in the target population only automatic steps are needed: the stemming of the texts of the target population, the word-stem profile counting to compute the vector $P(S)$, and finally the use of (9) to compute the estimated category proportions $P(D)$ of the target population. In detail, if corpus of texts is collected along time, like in our application, the evolution of the sentiment and opinion proportion along time can be depicted repeating the three steps above every for every selected time interval. Figure 3 shows a synthetic resume of the estimation strategy.

4 Analysis of Expo Milano 2015

The proposed method is now applied to the case of Expo Milano 2015 in order to evaluate how its web reputation evolved over time, before, during and after the event in both Italy and abroad.

Data were downloaded using the *twitteR* package (Gentry, 2012), which connects to Twitter official Search API. In details, to build the target population (i.e., tweets talking about Expo Milano Expo 2015), we run different queries using the hashtags: # expomilano2015, # expomilano, # expo2015milano, # expo2015 through Search API on tweets published in

both Italian and English. The output of this process is a corpus of texts with a list of 15 related attributes among which: tweet identification code, time, latitude and longitude, if the post is a retweet, number of likes and retweets received, conversation details such as users involved and users information such as name and source link. To avoid any data loss during the downloading process, our data were collected replicating the queries every day from the 17th of February 2015 to the 31st of December 2015 six times per day (1 a.m., 8 a.m., 12 a.m., 3 p.m., 6 p.m. and 9 p.m.).

The selection of the categories is based on the identification of the key components of the Expo reputation and it is driven by the analysis of available published sources about the exhibition. At a first level of analysis (i.e., sentiment analysis) we distinguish between five sentiment categories: Positive, Negative, and Neutral texts concerning Expo Milano 2015; Off-Topics texts, that are not related to Expo Milano 2015; and Advertise texts, that are specifically linked to sponsors and sponsoring related activities. This last category is introduced in this study in order to take into consideration the specificity of the event we are analysing. It is also particularly relevant from a managerial point of view, because it allows to analyse the visibility and the impact of these initiatives that are associated to a specific stakeholder category.

At a second level of analysis (i.e., opinion analysis), we identified 24 opinion categories. The Positive and Negative sentiment categories are segmented into ten categories reflecting different possible "drivers" of the sentiment. Six categories are aimed at capturing visitors' perceptions about their visit to the exposition site, ranging from the quality of food available, the site organization, the architecture of the pavilions, the costs of tickets and products, the technology innovation. Three categories are instead aimed at capturing the perception of the general public about administrative and organizational issues related to the development of the site, (bribes, building site delays, etc.), the sponsor selection and the quality of employment. In order to capture unexpected soaring discussions, some extra categories are added after a preliminary data exploration. An example are the categories concerning the feed-backs out-coming the manifestation against Expo, which took place in Milan the 1st of May. This label is defined due to the high number of tweets discussing about this topic during the days after.

After the category definition, the Random Hand-Coded set and the Modified Hand-Coded are built via manual tagging. In order to build the Random Hand-Coded set, 200 tweets per month are labelled into a sentiment and an opinion category. Then, we identify the rare categories, based on the relative weight of different categories in the Random Hand-Coded set. We

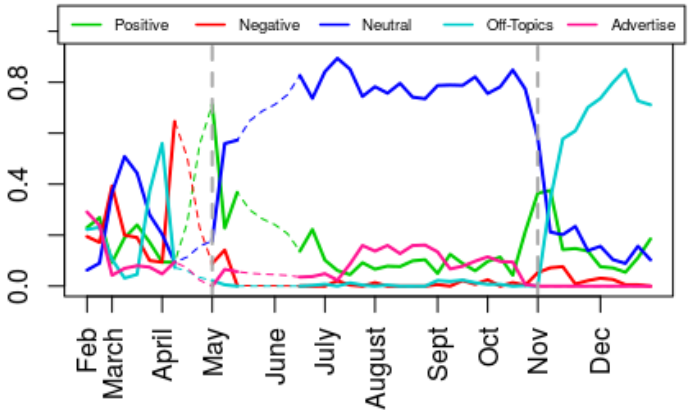
associate to each rare category a list of related words and adjectives in both languages (i.e. Category concerning the 1st May strike: “sciopero”; “1 Maggio”; “manifestazione”; “vandali” etc.. Category concerning the delay of the construction site and the corruption linked with the tender: “bustarella”; “corruzione”; “mafia” etc.). We perform keyword-based queries and, based on the results of these queries, we add on average, 50 texts per months to these rare categories. This process leads to the formulation of an over-sample Modified Hand-Coded set.

The stemming process is done in R using the *tm* package (Feinerer and Hornik, 2013) and the *SnowBallC* package (Bouchet-Valat, 2014). Particular attention in this phase is dedicated to the proper treatment of English words in the Italian data-set in order to avoid mis-identification errors. To this aim, the stemming process in the Italian data-set is done using both Italian and English words.

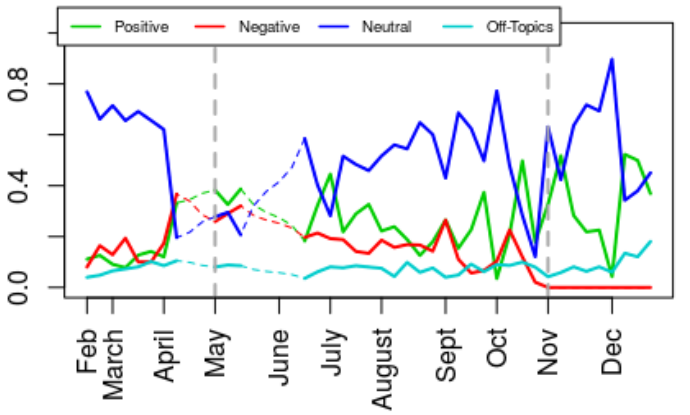
Classification performance evaluation shows that better performances are obtained if the analysis is segmented in three parts related to the life-cycle of the event: Pre-Expo (from 17th of February to the 30th of April), the Expo (1st of May to 31st of October) and Post-Expo (1st of November to 31st of December). Therefore, categories selection and final results estimation are conduced separately for these three time periods and separately for Italian and English. In the cross validation process, the estimation quality is measured using the (10). If the method is under-performing on some specific categories, those ones are tentatively merged into a more general opinion category. Classification performances are detailed in the Section 5

The final results represent the estimates of the sentiment and opinion categories, for the tweets published in Italian and in English in each selected week during the three periods of observation. In order to pinpoint the changes occurring to the Expo reputation, we plot the results in terms of category proportions (Figure 4).

Figure 4a highlights a clear dynamic in the Italian-speaking Twitter community, that changes over the three periods of observation. During the Pre-Expo weeks, the volume of different categories is comparable and there is not a prominent group. During the exhibition, the Neutral category is outdoing all the other categories. During the Post-Expo weeks, the Off topic category is outdoing the other categories. This dynamic has a methodological explanation and a practical interpretation. During the Pre-Expo and Post-Expo weeks, tweets about Expo are watered down compared to the ones posted during the Expo. Due to this behaviour, the classifier performs better during the exhibition, because the discussion about the Expo turned on after the opening and the data collected in that period are more representative



(a) Sentiment proportion: Italian Tweets



(b) Sentiment proportion: English Tweets

Figure 4: Sentiment analysis: Italian and English Tweets. Plot of the estimated proportions of the sentiment categories as a function of time. Raw proportions per month are displayed. Dotted lines represent interpolated missing data due to downloading problems.

of the problem of interest. Analysing the dynamics of the results further, two peaks can be registered during the Pre-Expo weeks. In April, the Figure 4a shows a negative sentiment peak, that was determined by recurrent discussions concerning delays in the preparation of the site, controversies about the role of volunteers and the starting of a inquiry that involved some relevant Expo managers with an accuse of corruption. The details about these dynamics are deepened in the opinion analysis results. Then, on the 1st of May, there is a high positive sentiment peak, that corresponds to the the Opening of the exhibition and the first Expo week. Instead, focusing on the Advertise category, no peaks but a gradual increase can be highlighted. This category indeed is growing in the central summer months of the Expo, such as August and September. In these months, many concerts took place in the Expo site. All these initiatives, like concerts and national dates, were strongly promoted by the official Expo Twitter account. Moreover, a massive advertising campaign was done to attract tourists during their holidays (in Italy usually from July to September).

It is also interesting to pinpoint the similarities and differences that emerge from the analysis by comparing English and Italian tweets (Figures 4a and 4b). First, the inversion of trends in the positive and negative sentiment, before and after the Opening of the exhibition, is confirmed even for the English data-set, however, compared to the Italian one, it is very slight. On the contrary, the Off-Topic category never slows down, suggesting that the crawling word "Expo" is collecting many differing topics in English.

In Figure 5, selected results of the opinion analysis are presented, to go deeper into the sentiment analysis. For what concern the Italian opinion categories, the highest peak is the discussion about the political management of the site, sponsor and structure, occurring in April, just before the opening. Then comparing different trends in the analysis of the Italian and the English data-sets, a few differences emerge. The Italian discourse is mainly focused on the discussion of administrative and organizational issues, whilst the English discourse concerns practical information and feedback strictly related to the visit.

The most attractive aspects of the Exposition for foreign visitors were design and food. Figure 5a and 5b show that the feed-backs about the architecture of the pavilions and the quality of food and beverage are proportionally higher in the English data set than in the Italian one. Along the whole period, visitors from abroad seem to appreciate the architecture of the pavilions and the quality of the food more than locals did.

Instead, the Italian public was more sensible to issues related to the site organization and management (as shown in Figure 5c), whereby negative

(a) Food and Beverage

(b) Architecture

(c) Organize and Manage Site

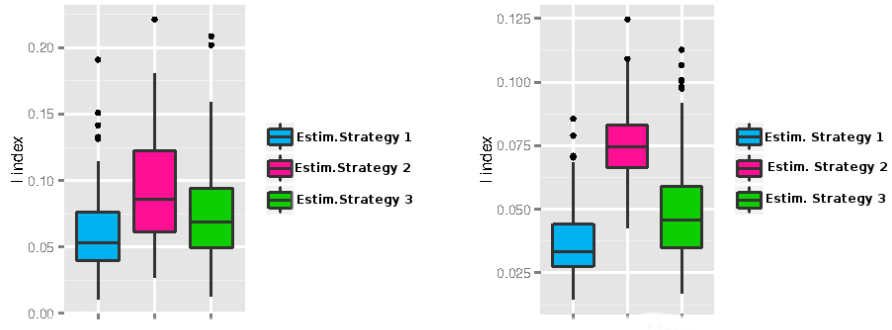
(d) Sponsor

Figure 5: Opinion analysis: Italian vs. English. Comparison between selected opinion categories in the Italian and in the English dataset. The plots represent the proportions as a function of time. In Figure 4 and Figure 5, the plots of the sentiment categories are displayed as function of time.

perceptions were proportionally higher in the Italian data-set, during the Pre-Expo and Post Expo weeks. This can be easily explained considering that administration and management of the site was somehow interrelated with different aspects of the political life of the country and, as a consequence, was widely discussed locally.

5 Comparison with other Estimation Strategies

In this section, firstly our classification scheme is compared with the Hopkins and King one (Hopkins and King, 2010). Secondly, its performances are estimated focusing on changing some interesting parameters. All the comparisons are evaluated via cross-validation using the index (10).



(a) Sentiment analysis.

(b) Opinion analysis.

Figure 6: Estimation Strategies evaluation: Sentiment and Opinion Analysis. Barplots of I index computed after $n = 100$ simulation of the three Estimation Strategy to estimate Sentiment and Opinion categories on the same test set.

Here below, a list of all the tested and compared estimation strategies (i.e., different combinations of a training set and a specific classification approach):

- Estimation Strategy 1 - Hopkins and King approach without Keywords-based reinforcement: cross-validation training sets are sub-sampled from the Random Hand-Coded set and the estimation strategy adopted is the method proposed by Hopkins and King;
- Estimation Strategy 2 - Hopkins ans King approach with Keywords-

based reinforcement: cross-validation training sets are sub-sampled from the Modified Hand-Coded set and the estimation strategy adopted is the method proposed by Hopkins and King;

- Estimation Strategy 3 - Rare But Not Least approach with Keywords-based reinforcement: cross-validation training sets are sub-sampled from the Random and the Modified Hand-Coded set and the estimation strategy adopted is the method described in 3

The cross-validation test sets are the same for all the experiments and are sub-sampled from the Random Hand-Coded set.

Figure 6 shows the index I cross-validation distribution of the Sentiment and Opinion categories estimation. Estimation strategy 2 is as expected the worst performing one. Estimation strategy 2 is indeed trained with a bias training set, without imposing any estimation correction. The most interesting comparison is instead between estimation strategy 1 and estimation strategy 3, because they are both providing unbiased estimates. Even though both estimation strategies asymptotically tends to the true categories proportion, simulations show in both sentiment and opinion analysis a slightly better performance of estimation strategy 1. Despite of that, estimation strategy 3 is giving the chance of depicting 30% more categories (i.e. seven ctegories) then estimation strategy 1, with just a small loss in terms of general performance.

The Rare-But-Not-Least classifier performances are influenced by the estimation of the two corrective coefficients A and B. A and B are computed using estimation of stems distributions from the population sampled with the keywords-bases reinforcement, i.e. the Modified Hand-Coded Set, and from an unbiased data-set. The unbiased data-set needs to fulfill the condition (5). In our experiment, it consists of the Random Hand-Coded set, which surely fulfills the hypothesis. In Figure 7, the Rbnl classifier performance, computed using (10), are monitored tuning the dimension of the Random Hand-Coded set L. As expected, increasing L leads to a better estimation of the A and B coefficients, and consequently to a better performance of the classifier. Indeed, a more precise correction of the Modified Hand-Coded set distortion is measured when the A and B coefficients are computed on a richer unbiased data set, which is better representing the target population.

Figure 7: Estimation Strategy 3 Performance. Plot of I index of Estimation Strategy 3 as a function of L, dimension of dataset on which the corrective coefficient A and B are built. The values of I index refers to sentiment analysis.

6 Conclusions

In the era of the Internet and social media, information and communication technologies (ICT) have facilitated enormously the process of generation and diffusion of information, offering users the possibility to both retrieve contents from plenty of different sources and create and share their own contents with other users. In this way, a vast amount of data has become readily available to decision-makers in almost any field of human activity, and they started to look at different social media (such as reviews, forum discussions, blogs, micro-blogs, Twitter, Facebook ecc.) as a potential means for understanding the opinions of their clients, users, and stakeholders in general (Cheng, Chiang, and Storey, 2012).

However, for exploiting opinionated data derived from different social media to extract relevant information, academics and practitioners have to deal with different elements of complexity that are inherent to the characteristics of these data - they are indeed very heterogeneous, in terms of contents and formats and are generated (and change) very frequently. In this context, this paper has addressed one specific element of complexity, i.e. the presence of rare text categories in opinionated data. This situation, which is frequent in practice, is pretty critical from a managerial perspective.

Indeed, if decision makers aim to use social media data to support decision making processes, they cannot overlook rare categories, just because they are less frequent than the others. This could lead to the inability of capturing emerging phenomena or punctual critical issues. Still form a methodological perspective, in presence of rare categories, there is a non-null probability of not gathering any text belonging to these categories in the training set, with the risk of losing some relevant pieces of information. To deal with this problem, the proposed aggregated classification scheme improves the ability of sentiment and opinion analysis to capture rare categories. Firstly, the new classification scheme save precious time in building the training set, by sampling and manually tagging only the texts required to well represent an interesting category. Secondly, the sampling correction and rescale ensures unbiased final estimations of either rare and not rare categories proportions with a global estimation error measured via cross-validation comparable with the one provided by the basic aggregate supervised classifier

With respect to the specific application, the sentiment analysis supports the achievement of a general understanding about the event reputation on the web. In the English data-set, no drastic difference between the positive and the negative trends are registered, while in the Italian data-set there is a trend inversion between Pre and Post Expo. Enthusiasm about the mega event takes the place of the initial distrust and criticism. The opinion analysis supports a more in-depths comprehension of the "why" and "how" the sentiment dynamics take place. The selected time interval is thin enough to depict events and time changes. From corruption to concerts, from sponsor critiques to the 1st of May strike, the opinion analysis allows to trace different drivers of stakeholders' perceptions.

An instant further development is to test the new classifier scheme on a real time monitoring problem, such as the political election campaign. New trends and topics appearing during the study can be added to the training set with the keywords reinforcement sampling method. This solution allows tracking real time the new and old categories without rebuilt the training set from the beginning. Another method development can be applied to the training set adaptation in a cross domain learning based study. The training set from the source domain can be integrated with text from the target domain, to better perform on the target test set. The algorithm will be implemented as a running option of the next release of R package ReadMe.

Appendix A

Proof of Theorem (1).

Consider the result (1), the demonstration in the simplest case is here shown. If $K = 1$ and $J = 2$: the possible word stem profiles are $S_1 = 1$ and $S_2 = 0$ and only two categories exist d_1, d_2. Note that, the words stem profile in the random training and cased control sampled training are the same 2^K. The difference lies in their probabilities. In the following counts the superscript H which describes the hand-coded training set is omitted. Consider the matrices and applying the Bayes theorem:

$$A_1 P^{MODH}(S = 1|D = D_1)B_1 =$$
$$= P^{MODH}(S = 1|D = D_1)\frac{P^{RH}(S = 1)}{P^{MODH}(S = 1)}B_1 =$$
$$= PMODH(D = D_1|S = 1)\frac{P^{MODH}(S = 1)}{P^{MODH}(D = D_1)}\frac{P^{RH}(S = 1)}{P^{MODH}(S = 1)}B_1$$

Using the hypothesis of the cased control sampling $P(D|S) = P(d|s)$:

$$P^{MODH}(D = D_1|S = 1)\frac{P^{MODH}(S = 1)}{P^{MODH}(D = D_1)}\frac{P^{RH}(S = 1)}{P^{MODH}(S = 1)}B_1 =$$
$$= P^{RH}(D = D_1|S = 1)\frac{P^{RH}(S = 1)}{P^{MODH}(D = D_1)}B_1 =$$
$$= P^{RH}(D = D_1|S = 1)P^{RH}(S = 1)\left[P^{MODH}(D = D_1)\right]^{-1}$$
$$\left[P^{MODH}(D = D_1|S = 1)\frac{P^{RH}(S = 1)}{P^{MODH}(D = D_1)} + P^{MODH}(D = D_1|S = 0)\frac{P^{RH}(S = 0)}{P^{MODH}(D = D_1)}\right]^{-1}$$

Using the theorem of the total probability:

$$\frac{P^{RH}(D = D_1|S = 1)P^{RH}(S = 1)}{P^{RH}(D = D_1|S = 1)P^{RH}(S = 1) + P^{RH}(D = D_1|S = 0)P^{RH}(S = 0)} =$$
$$= P^{RH}(D = D_1|S = 1)\frac{P^{RH}(S = 1)}{P^{RH}(D = D_1)} =$$
$$= P^{RH}(S = 1|D = D_1)$$

The following final equation comes from the model Hypothesis (2).

$$A_1 P^{MODH}(S = 1|D = D_1)B_1 = P^{RH}(S = 1|D = D_1) = P(S = 1|D = D_1) \quad (11)$$

The result is easily extended to the more complex case where $K > 1$.

References

Berger, Adam L., Vincent J. Della Pietra, and Stephen A. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, Vol.22,1 39-71

Blei, David (2012) Probabilistic topic models. *Communications of the ACM* 55(4):7784.

Blei, David M., and John D. Lafferty (2007) A correlated topic model of science. *The Annals of Applied Statistics* (2007): 17-35.

Blei, David M and Ng, Andrew Y et al. (2003). Latent Dirichlet Allocation. *the Journal of machine Learning research*, Vol.3, JMLR.org

Bouchet-Valat, M (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. *R package version 0.5. 1, URL http://CRAN. R-project. org/package= SnowballC*

Breimen, L and Friedman, JH and Olshen, RA and Stone, CJ (1984) Classficatiun and Regression Trees. *Wadsworth, Belmont*

Breslow, Norman E (1996) Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association, Taylor & Francis Group* Vol. 91,433

Ceron, Curini, and Iacus (2013) Unsing Social media to forecase electoral results:a review of the state of the art. *Italian Journal of Applied Statistic* Vol. 25,3

Ceron, Andrea, Luigi Curini, and Stefano Maria Iacus(2016) iSA: a fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Science*

Chen, H., Chiang, R. H., and Storey, V. C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly* 36(4): 1165-1188

Choi, Dongjin, and Pankoo Kim (2013) Sentiment analysis for tracking breaking events: a case study on twitter. *Asian Conference on Intelligent Information and Database Systems. Springer Berlin Heidelberg*

Corallo, Angelo, Fortunato, Laura, Matera, Marco, et al. (2015) Sentiment Analysis for Government: An Optimized Approach *Machine Learning and Data Mining in Pattern Recognition, MLDM Book Series: Lecture Notes in Artificial Intelligence* Vol. 9166: 98-11

Das, Sanjiv, and Mike Chen(2001) Yahoo! for Amazon: Extracting market sentiment from stock message boards. *Proceedings of the Asia Pacific finance association annual conference (APFA)* Vol. 35. 2001.

Nirmala, C. R., G. M. Roopa, and KR Naveen Kumar (2015) Twitter data analysis for unemployment crisis. *International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE*

Dave, Kushal, Steve Lawrence, and David M. Pennock. (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web. ACM*

da Silva, Hruschka and Hruschka Jr (2014) Tweet sentiment analysis with classifier ensembles *DECISION SUPPORT SYSTEMS* Volume: 66 Pages: 170-179 Published: Oct. 2014

Duda, Richard O., and Peter E. Hart. (1973) Pattern classification and scene analysis. *New York: Wiley, Vol 3.*

Erosheva, Elena, Stephen Fienberg, and John Lafferty Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004): 5220-5227.

Expo 2015 S.p.a. *http://www.expo2015.org/*

Feinerer, Ingo and Hornik, K (2013) *tm: Text Mining Package. R package version 0.5-9.1*

Gentry, Jeff (2012) twitteR: R based Twitter client. *R package version 0.99* Vol. 19

Go, Alec and Bhayani, Richa and Huang, Lei (2009) Twitter sentiment classification using distant supervision *CS224N Project Report, Stanford* Vol. 1

Grimmer and Brandon (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*

Hand, David J. (2006) Classifier technology and the illusion of progress. *Statistical Science* 21(1):115.

Hopkins, Daniel J and King, Gary(2010) A method of automated nonparametric content analysis for social science. *American Journal of Political Science* Vol. 54, 1, Wiley Online Library

Hopkins, Daniel and King, Gary and Knowles, Matthew and Melendez, Steven(2010) ReadMe: Software for automated content analysis *Institute for Quantitative Social Science*

Joachims, Thorsten. (1998) Text categorization with support vector machines: Learning with many relevant features. *The Annals of Applied Statistics* European conference on machine learning. Springer Berlin Heidelberg

King, Gary and Zeng, Langche (2001) Logistic regression in rare events data *Political analysis* Vol 9, 2, SPM-PMSAPSA

Laver Michael, Kenneth Benoit, and John Garry (2003) Extracting policy positions from political texts using words as data. *American Political Science* Review 97.02, 311-331

Lowe, Will (2008) Understanding wordscores. *Political Analysis* 16.4 (2008): 356-371.

Martin, Lanny W., and Georg Vanberg (2008) A robust transformation procedure for interpreting political text. *Political Analysis* 16.1 (2008): 93-100.

Mahalakshmi and Sivasankar Cross Domain Sentiment Analysis Using Different Machine Learning Techniques *5th International Conference on Fuzzy and Neuro Computing (FANCCO)* 17-19, 2015

Monroe, Burt L., and Ko Maeda. (2004) Talks cheap: Text-based estimation of rhetorical ideal-points. *Annual meeting of the Society for Political Methodology.*

Mudinas, Andrius, Dell Zhang, and Mark Levene (2012) Combining lexicon and learning based approaches for concept-level sentiment analysis. *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining.* ACM, 2012.

Mukherjee, Subhabrata, and Pushpak Bhattacharyya (2013) Sentiment analysis: A literature survey. *arXiv* 1304.4520

Pang and Lee (2008) Opinion mining and sentiment analysis. *Foundations and trends in information retrieval* 2.1-2: 1-135.

Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar (2002) Thumbs up?: sentiment classification using machine learning techniques

Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Vol.10, Association for Computational Linguistics

Rao et al. (2014) Sentiment topic models for social emotion mining. *Information Sciences* 266 (2014): 90-100.

Rao, Yanghui; Lei, Jingsheng; Liu Wenyin; et al. (2014) Building emotional dictionary for sentiment analysis of online news *World Wide Web-Internet and Web Information System* Vol. 17,4: 723-742

Ribeiro, Filipe N., et al. (2016) SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5.1 (2016): 1-29.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi (2016) A model of text for experimentation in the social sciences. *Journal of the American Statistical Association* (2016): 1-49.

Salter-Townshend, Michael, and Thomas Brendan Murphy (2014) Mixtures of biased sentiment analysers. *Advances in Data Analysis and Classification* 8.1 (2014): 85-103.

Slapin, Jonathan B, and Proksch, Sven-Oliver (2008) A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, Vol 52,3. Wiley Online Library

Stone, Dunphy, Smith and Ogilvie (1968) The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, Vol 8.1 113-116

Taboada, Maite, et al. (2011) Lexicon-based methods for sentiment analysis. *Computational linguistics* 37.2 (2011): 267-307.

Tian, Feng, et al. (2016). A topic sentence-based instance transfer method for imbalanced sentiment classification of Chinese product reviews. *Electronic Commerce Research and Applications* 16 (2016): 66-76.

Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath (2016) Classification of sentiment reviews using n-gram machine learning approach *EXPERT SYSTEMS WITH APPLICATIONS* Volume: 57 Pages: 117-126

Zhao, Hua, et al. (2016) A teaching evaluation method based on sentiment classification. *International Journal of Computing Science and Mathematics* 7.1 (2016): 54-62.

Zhou, Zhixin, Xiuzhen Zhang, and Mark Sanderson (2014) Sentiment analysis on twitter through topic-based lexicon expansion. *Australasian Database Conference.* Springer International Publishing, 2014.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

01/2017 Riccobelli, D.; Ciarletta, P.
Rayleigh-Taylor instability in soft elastic layers

58/2016 Antonietti, P. F.; Bruggi, M. ; Scacchi, S.; Verani, M.
On the Virtual Element Method for Topology Optimization on polygonal meshes: a numerical study

56/2016 Guerciotti, B.; Vergara, C.; Ippolito, S.; Quarteroni, A.; Antona, C.; Scrofani, R.
A computational fluid-structure interaction analysis of coronary Y-grafts

57/2016 Bassi, C.; Abbà, A.; Bonaventura, L.; Valdettaro, L.
Large Eddy Simulation of gravity currents with a high order DG method

55/2016 Antonietti, P. F.; Facciola' C.; Russo A.; Verani M.;
Discontinuous Galerkin approximation of flows in fractured porous media on polytopic grids

54/2016 Vergara, C.; Le Van, D.; Quadrio, M.; Formaggia, L.; Domanin, M.
Large Eddy Simulations of blood dynamics in abdominal aortic aneurysms

52/2016 Paolucci, R.; Evangelista, L.; Mazzieri, I.; Schiappapietra, E.
The 3D Numerical Simulation of Near-Source Ground Motion during the Marsica Earthquake, Central Italy, 100 years later

53/2016 Antonietti, P. F.; Manzini, G.; Verani, M.
The fully nonconforming Virtual Element method for biharmonic problems

51/2016 Guzzetti, S.; Perotto, S.; Veneziani, A.
Hierarchical Model Reduction for Incompressible Flows in Cylindrical Domains: The Axisymmetric Case

48/2016 Scardulla, S.; Pasta, S.; D'Acquisto, L.; Sciacca, S.; Agnese, V.; Vergara, C.; Quarteroni, A.; C
Shear Stress Alterations in the Celiac Trunk of Patients with Continuous-Flow Left Ventricular Assist Device by In-Silico and In-Vitro Flow Analysis