2011 11th IEEE International Conference on Data Mining Workshops

Detecting General Opinions from Customer Surveys

Evgeny A. Stepanov, Giuseppe Riccardi
Department of Information Engineering and Computer Science
University of Trento
Trento, Italy
{*stepanov,riccardi*}*@disi.unitn.it*

***Abstract*—Questionnaire-based surveys and on-line product reviews resemble each other in that they both have user comments and satisfaction ratings. Since a comment might be a general opinion about a product or only one or a set of its attributes, in which case the text might not reflect the rating; surveys and reviews share the problem of pairing free-text comments with these ratings. To train accurate models for automatic evaluation of products from free-text, it is important to distinguish these two kinds of opinions. In this paper we present experiments on detecting general opinions that target a product as a whole; thus, reflect the user sentiments better. The task is different from subjectivity detection, since the goal is to detect generality of an opinion regardless of the rest of the documents being opinionated or not. The task complements feature-based opinion analysis and opinion polarity classification, since it can be applied as a preceding step to both tasks. We show that when used as a classification feature user ratings are not useful in the general opinion detection task. However, they are effective in predicting the polarity of a comment once it is identified as a general opinion.**

***Keywords*-Opinion mining, Sentiment analysis, Classification.**

I. Introduction

Questionnaire-based surveys are traditional tools designed to collect information about user (or customer) experience, opinions, etc. in disciplines such as social and marketing sciences. Questionnaires include sets of pointed questions addressing specific or general properties of a product, services or behavior. Most of the time such questionnaires include free-text forms the surveyee may (or may not) use to elaborate in natural language his/her opinion, augment the scope of the survey, or improve the quality of the general-purpose structure of the questionnaire. The analysis of the closed-answer part is relatively easy, since it is a structured data and several multivariate analysis techniques exist; however, it is limited in scope. The analysis of the open-answer part is not as easy, but it might contain opinions within and outside of the scope of the questionnaire, objective problem statements, suggestions, etc.. Detection of opinions and their polarities and the domain topic classification are important aspects of the analysis of such open-ended user comments.

The structure of questionnaire-based surveys resembles the structure of on-line product reviews one can find on the Internet. These reviews generally provide some kind of rating expressing users' satisfaction with a product. With respect to the method of collecting user opinions about a product, web-sites could be grouped as providing a general rating only or also ratings for different attributes of the product. Among the most popular web-sites one can compare IMDb.com to Yahoo! Movies and Amazon.com to Epinions.com to see the examples of these. The latter (Yahoo! Movies and Epinions.com) clearly bear more resemblance to the traditional questionnaire-based surveys.

The problem both surveys and on-line reviews share is pairing of a free-text comment with a user rating. In the field of opinion mining and sentiment analysis, it is generally assumed that these user ratings are good indicators of polarity of a document; and such signal is used to train and evaluate tasks such as opinion polarity classification. Often, the practice is to remove documents that belong to the ambiguous "middle" part of the rating scale.

However, the cases when the user score is not reflecting the true "rating" of the text are known to exist [1]. User may choose to use rating and the comment for different purposes. For example, to use rating to reflect their true opinion about the object, and a comment to provide additional information or explain their reasons for the provided rating. Moreover, user might elaborate only about negative aspects while providing a high score, which will pair a negative text with a positive rating. Thus, automatic pairing of a free-text and a rating yields inaccurate model for classification, and therefore, when such a model is used, a product or service might not receive a fair evaluation from textual data only (e.g. from blogs, etc.). A possible solution to this problem is to identify comments (or their parts) that target a product as a whole rather than one of its attributes, that is *general opinions*, and then decide on the polarity of a comment.

The task is different from subjectivity detection [2], where the objective is to detect opinionated text. In general opinion detection task, the objective is to detect the generality of an opinion regardless of the subjectivity of the rest of the documents. The task complements feature-based opinion analysis [3], since instead of focusing on extraction of text fragments on specific attributes of a product and detection of their polarities, it only focuses on detection of text fragments targeting a product or service as a whole. The general opinion detection also can be applied as a first step to opinion polarity (sentiment) classification [4], where the goal

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.63

is to assign a satisfaction degree (either positive or negative, or a satisfaction rating) to a text, and which assumes that this text is opinionated.

In this paper we present a case study on telecom customer care surveys including directed questions related to products and services as well as free-text user comments. We describe a supervised machine learning approach to the general opinion detection and polarity classification problems on the real document distribution. The objective is to answer questions such as: (1) Are user provided ratings useful to identify documents bearing general opinions? (2) Are user ratings effective in predicting opinion polarity of their free-text comments?

The rest of the paper is organized as follows. Section II gives a short summary of the related work. Section III describes the methodology for detecting general opinions and polarity classification. Then, Section IV describes the nature of the analyzed document collection and provides details on properties of the documents and features. Section V presents experiments and results on General Opinion Detection and Polarity Classification using various training settings. Section VI provides concluding remarks.

II. Related Work

Following [5], we group the work done in Opinion Mining into Sentiment Classification and Feature-based Opinion Mining. Sentiment Classification is further divided into Opinion Detection and Sentiment Polarity Classification.

A significant amount of research has been focused on opinion detection on various levels of document granularity. The finer the level of granularity, the more complex is the task [6], [7]. Opinion detection on the document level is closely related to genre classification, and the problem is usually cast so [8].

Sentiment polarity classification has attracted a significant amount of attention as well. The problem is usually defined either as a binary classification to identify a text fragment as either positive or negative; or as a regression to predict user rating. Similar to opinion detection, it is approached on various levels of granularity: words, phrases, sentences, or whole documents. The general practice in the field is to use the user rating indicator as a supervision label for the classification: e.g. [9] used numeric satisfaction scores from 1 to 4 as classification classes; [10], [4] and [11] used user ratings for an automatic selection of negative and positive examples. It was shown that separation of factual information from subjective one has a positive effect on polarity classification [11], [12].

The closely related task is feature (aspect)-based opinion mining (e.g. [3] and [13]). Given a document, the task is usually defined as assignment of a polarities to a set of product attributes (including the whole) with respect to that document. The features set is either predefined or automatically extracted from the data. The result of feature-based opinion mining is the summary over a set of documents in a form of the attribute set with related polarity counts.

Even though Named-Entity Extraction covers identification of a target of an opinion (Item Extraction [14]) to some degree, the task is often overlooked. However, since even in direct questionnaires and reviews an expressed opinion might have a different target (e.g. an actor in a movie review [14]), it is critical.

There are two main approaches to the mentioned tasks. The first is to use supervised machine learning techniques to learn subjectivity or polarity classification from manually or automatically labeled set of documents. The second is to use manually or automatically created resources (e.g. sentiment dictionaries), and starting from a low level classification (word) to compute polarity of a document with respect to some function. In this paper we use supervised machine learning for general opinion detection and polarity classification from manually annotated data.

III. Methodology

An opinion can be expressed about an entity as a whole or just about one or a set of its attributes. When one desires to evaluate an entity as a whole, it is important to distinguish the granularity of expressed opinion as well as its polarity. The goal of the experiments is to assess the complexity of the general opinion detection with subsequent polarity classification. We compare two factors that affect the task.

(1) Very often freely available data sets represent ideal case, i.e. the distribution of categories is balanced. Unfortunately, this is rarely the case in practical scenarios. An unbalanced distribution of categories in a data set usually results in a category-specific performance being dependent on the number of samples of that category in the data set. In this paper we contrast performances of the models trained on balanced data sets and data sets with natural distribution. Random under-sampling is used for balancing the data set.

(2) The fact that user ratings do not always reflect the polarity of a free-text comment, as well as the availability of manually labeled data (described in the following Section IV) allow us to asses these user ratings as a feature, rather that as a supervision signal for both general opinion detection and polarity classification tasks.

A. General Opinion Detection

Generally, the goal of *Opinion Detection* is to classify whether a unit of classification is factual or expresses an opinion on a target entity (i.e. person, product, etc.). The task can be performed using different granularities: a unit of classification might be a word, a phrase, a sentence, or a document. A unit categorized as subjective can be further classified as having negative or positive polarity, which is the task of *Polarity Classification*. Opinion detection is considered to be a harder problem than topic classification and opinion polarity classification [7].

Our goal, however, is to detect *general* opinions, regardless whether the rest of the documents is opinionated or not. Even though a comment expressing an opinion on an attribute of a product or service can often be seen as more factual than a general opinion (since, depending on an attribute, it is possible to evaluate the validity of statements), subjective vs. factual classification is not viable. Therefore, we experiment with different settings: (1) One-against-all binary classification; and (2) multi-class multi-label classification. In the first setting general opinions are contrasted to the rest of documents, whereas in the second one they are contrasted to subgroups such as opinions about attributes of a product (usually problem reports), suggestions, etc..

B. Polarity Classification

In Section V we present the results of experiments on Polarity Classification, where general opinion category documents were first extracted to form a separate data set. We experiment on a data set with natural distribution of labels, as well as use random under-sampling to balance the data. The goal is to asses the effectiveness of user ratings in predicting the polarity of free-text comments.

C. Classifiers

For each task two state-of-the art classifiers were used for the experiments: (1) BoosTexter [15], a boosting-based machine learning program for text classification; and (2) SVM^{light} [16], a popular Support Vector Machines [17] implementation frequently used for text categorization.

The idea of Boosting family of algorithms is to combine a number of weak classifiers into a strong one. BoosTexter learns a weak classifier each iteration, and the classification model (i.e. strong classifier) is updated. There is no build-it procedure to terminate the training process, like in the case of SVM^{light}. Consequently, for BoosTexter we evaluate two models, selected with respect to the following heuristics: (1) to keep the fixed number of iterations equal to the training data size, and (2) to choose the model with the least training error among these iterations (algorithm is optimizing training error). Neither of the two heuristics guarantees the optimal model; however, they provide reasonable estimates.

D. Experimental Set-Up

All the experiments in Section V are carried out on the training set of the corpus described in Section IV. For classification training and testing stratified sampling with 5-fold cross-validation (further *CV5*) is used.

IV. Data Set

The document collection used in the experiments through out the paper is Telecom Italia Customer Care Surveys (further referred as TI Data Set) from 3 consecutive years (2 surveys a year). A document in the survey is a combination of a structured questionnaire (question – answers pairs) and a user's comment. The data set constitutes a subset of a larger collection, selected such that a document contains both a questionnaire and a comment. The domain of the documents is products and services provided by telecom operators. The language of all the documents is Italian.

A. Questionnaire

The questionnaire for the direct elicitation of opinions about products (software) consists of a set of 27 questions requiring either multiple choice answers[1] or answers on a 1 to 10 scale[2]. The questions are grouped with respect to the information or judgment they elicit as following: general user information related to the use of a product (3), functionality of a product on aspects such as a quality of visualization, an interface with other software, etc. (11), product's performance (e.g. speed) (2), quality of technical support for a product (7), related training and documentation (3), and general satisfaction with a product and associated services, i.e. a global satisfaction (1).

Most of the questions are of the form "*How do you evaluate the simplicity of the information visualization of our system?*". The question for eliciting the *global satisfaction score* is different: "*Thinking about the services you receive while using our system, consider functionality, technical performance, support, and training and documentation for learning the system. To what degree you are satisfied?*".

B. User Comments

In the open-answer part user is free to provide any feedback. This free-text section of the documents is a noisy textual data, frequently containing spelling errors and incomplete and ungrammatical sentences. Moreover, a significant portion of comments is affected by orthographic variations and other user generated lexical variations such as emphatic expressions: *mooooolta* (regular: molta - much), *cambiateloooooooo* (regular: cambiatelo - change it), *Formazione ricevuta NULLA* (Received NO training).

The length and style of user comments vary. The length ranges from a single word to a more than 300 words. The style, on the other hand, varies from a well formed paragraph to a list of points. Given all these properties, the data fall under the umbrella of the noisy text, which limits the number of linguistic features that could be exploited for the classification.

1) User Comment Annotation: The free-text section of the survey is manually annotated by domain experts into 15 categories with respect to the topic of user comments: 11 labels for different product aspects users were reporting *problems* about; 1 label for the *suggestions*; 2 labels for

[1]Multiple choice questions elicit objective factual information; thus, they are excluded from the analysis.

[2]The questions skipped by the users are scored as '0'. In whole document collection their number is not very high, and proportionally distributed among categories of interest.

Table I
DISTRIBUTION OF CATEGORIES ON THE META-LEVEL OF TI DATA SET

Category	Training		Testing	
Problems	2324	57.9%	389	67.4%
Suggestions	926	23.1%	67	11.6%
Gen. Opinions	266	6.6%	19	3.3%
Other	321	8.0%	41	7.1%
Problem & Gen. Opinion	40	1.0%	20	3.5%
Suggestion & Gen. Opinion	10	0.3%	1	0.2%
Problem & Suggestion	125	3.1%	39	6.8%
Problem & Other	0	0.0%	1	0.2%
Total	4012	100.0%	577	100%

the *general opinions*: negative and positive; and 1 label for everything else (i.e. *other*). There is no restriction on the number of labels a comment can be annotated with. However, most of the comments have a single label (85%).

A pair of user comments together with their English translations is given in the minipage. Even though both comments have a negative user rating, only the first one is judged by the expert to be a general opinion about the product; whereas the second one is judged to be a problem statement. The object of opinion of the former comment is the system, whereas in the latter it is its performance.

General Negative Opinion (user raiting 4):
Il sistema $\mathbf{product_i}$ va bene per la clientela Consumer e Business di piccole dimensioni. E' del tutto inadeguato per la clientela Corporate.
The system $\mathbf{product_i}$ *is good for small size Consumer and Business customers. It is totally inadequate for Corporate customers.*

Performance Problem (user rating 3):
Il problema maggiore e' sicuramente la lentezza del sistema in generale nell' effettuare transazioni.
The biggest problem, certainly, is the general slowness of the system in making transactions.

2) Training and Test Sets: Since data set consists of surveys collected in consecutive time periods, the partitioning of data into training and test is straightforward: chronologically the latest survey naturally constitutes the test set, whereas the rest was combined to form a training set.[3] The test set remains completely unseen and unanalyzed throughout the experiments.

3) Organization of Labels: The labels provided by the expert are grouped into natural meta-categories: *Problems*, *Suggestions*, *(General) Opinions*, and *Others*. The details on label distribution are presented on Table I. This 4-category label set is used for the experiments and further will be referred as a 4-way meta-level. Even though the label hierarchy is preserved, due to the data sparseness experiments are mainly done on the meta-level only.

C. Data Set Analysis

A document in TI Data Set consists of two parts: question elicited ratings and a free-text user comment. For the purposes of this work the focus is only on the global satisfaction score (GSS) – our user rating equivalent. As it was already mentioned, user ratings, such as *global satisfaction score*, are assumed to be good indicators of the document polarity. In our case, each user rating is paired with the expert judgment on the comment's being a general opinion or not; in case of general opinions, also with the judgment about its polarity.

[3]The latest survey was received by the authors in later period as well.

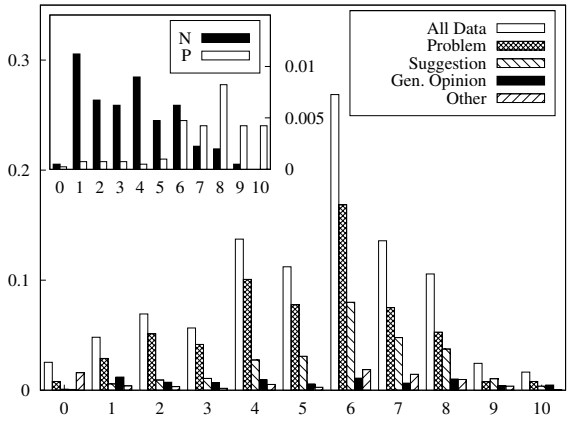

Figure 1. Global Satisfaction Score distribution for the whole TI Training Set and per class on the meta-level; and the distribution for General **P**ositive and **N**egative Opinions within General Opinion category (upper-left plot). (x-axis: Global Satisfaction Score; y-axis: Frequency (%)).

Figure 1 plots the distribution of global satisfaction scores for all training data and each label separately. The distributions do not show any modality. However, if we select a corpus from the entire set of documents, where user comments were annotated as general negative or positive opinions by the domain expert, the distribution is bimodal (see upper-left plot on Figure 1). Thus, it is possible to predict document polarity in an unsupervised manner after it was identified as a general opinion. However, based on the distribution, the data will also contain noise. In the following section we describe experiments using GSS as a feature for general opinion detection and for polarity classification.

V. EXPERIMENTAL RESULTS

Prior to proceeding with the main set of experiments on general opinion detection and polarity classification, a preliminary set of experiments was performed to assess the TI Data Set as well as the effectiveness of our processing and classification on a publicly available data set, the Movie Reviews [11].

A. Movie Reviews Baseline

Polarity data set 2.0 of [11] (Movie Reviews further on) is used to position BoosTexter unigram (bag-of-words) model with respect to SVM and Naïve Bayes classifiers used by the authors. The data set consists of 1000 negative and 1000 positive documents of IMDb movie reviews. The same 10-fold cross-validation split is used, 9 folds are used for training, 1 for test.

Table II
USER COMMENT POLARITY CLASSIFICATION ON TI DATA SET: AVERAGE ACCURACIES ON 10 FOLD CROSS-VALIDATION; AUTOMATIC SPLIT INTO NEGATIVE AND POSITIVE POLARITY USING USER RATINGS (GSS).

Features	Classifier	Accuracy
Unigram	SVM	**0.642**
	BoosTexter (H_1)	0.636
	BoosTexter (H_2)	0.640

To compare our data set to [11], we use random under-sampling to select 1000 documents with negative and 1000 documents with positive user ratings (GSS). On a 10 point scale, 4 and 7 inclusive thresholds are used, i.e. comments with GSS 5 and 6 (also unrated ones – 0s) are removed. The experimental settings for both data sets are identical: only the user comment part of the documents is used.

On the Movie Reviews data set, the average accuracy of the first BoosTexter heuristic (see Section III) on the 10-fold cross-validation is 0.831, and of the second is 0.807, compared to 0.872 and 0.828 reported by [11] for SVM and Naïve Bayes classifiers respectively. Our SVM model has 0.863 accuracy: we attribute the difference to the unigram pre-processing step described in [4], which we do not use.

Table II presents classifier performances in terms of average accuracy on 10 fold cross-validation; the same 2 heuristics are used for BoosTexter (the performance is not guaranteed to be optimal). We do not observe any significant performance differences between SVM and either of the BoosTexter heuristics.

In the following sections we describe experiments for each step in the classification process – General Opinion Detection and Opinion Polarity Classification tasks. For BoosTexter, throughout the rest of the experiments we utilize the second heuristic: train the classification models such that maximum number of iterations is equal to the size of the training set, and select the one with the lowest training error.

B. General Opinion Detection Task

Our experiments on General Opinion Detection are performed on a document level using different number of classes: (1) One-against-all binary classification; (2) 4-way classification into General Opinions, Problems, Suggestions, and Others, i.e. meta-level. For each setting the models are trained using all available data, i.e. natural distribution; as well as, a data set balanced using random under-sampling.

The data set for the balanced 4-way classification consists of 1,000 documents: randomly selected 250 document for each category.[4] For the balanced binary classification, the data set is 500 documents: 250 of the general opinion category, and the remaining 250 were randomly chosen among the other 3 meta-level categories, such that they are equally represented. In all the settings 5-fold cross-validation is applied.

[4]Multi-label documents were removed.

Since our goal is to detect a single category rather than all 4, for the SVM we apply the standard one-vs-one multi-class training strategy and train three binary General Opinion-versus-X classifiers, where X is either Problem, Suggestion, or Other. The final decision on General Opinion category membership is taken using majority vote of these 3 classifiers. Even though more advanced techniques exist (e.g. [18]), Majority Voting is a reasonable choice for a balanced data set.

The features used for classification are global satisfaction score (GSS) from the questionnaire and n-grams derived from user comments. Tables III and IV present 5-fold cross-validation results of using the features alone and in combination for training with natural and balanced distribution of labels, respectively.

For the binary classification setting one can observe that both classifiers fail to recall any general opinion document, when trained using only GSS feature. For the unigram only case, BoosTexter performs significantly better than SVM (Paired Two-Tail T-Test: $t(4) = 2.85, p < 0.05$). SVM trained on both features has zero recall as well; whereas BoosTexter achieves the best average F-measure (0.307). However, it is not significant in comparison to the unigram-only classifier with F-measure of 0.288 (Paired Two-Tail T-Test: $t(4) = -1.09, p > 0.05$).

Since our data set is multi-class and multi-label (4.4%) and a multi-class classification with SVM^{light} requires either a binary separation or treating multi-label documents as a separate class. SVM^{light} was not used in a 4-way classification on the data with natural distribution of labels (since either the setting would not be 4-way any longer, or the the distribution would cease to be natural.)

The same effect was observed for BoosTexter trained using the same features (GSS + unigrams) for a 4-way classification (i.e. the meta-level of the TI Data Set shown in Table I). GSS only classifier does not recall any general opinion document (see Table III). While the performance of the Unigram + GSS classifier is the highest (0.341), the difference is not significant in comparison to the classifier trained on unigram only features (0.320): (Paired Two-Tail T-Test: $t(4) = -1.34, p > 0.05$)

As it was predicted, GSS had no significant effect on General Opinion Detection; thus, unigrams or other features extracted from user comments could be used. Another observation is that higher F-measures on multi-class classification are not significant in comparison to the binary classification on a data set with natural distribution (Two-Tail T-Test assuming equal variances: $t(8) = -0.82, p > 0.05$ for unigram only classifier, and $t(8) = -0.86, p > 0.05$ for the classifier with feature combination).

Comparing results in Table IV with the results in Table III one can see that balancing the data set looks very effective when we consider the results of cross-validation. Even though, for BoosTexter we observe a positive effect

Table III
GENERAL OPINION DETECTION TASK: BINARY AND 4-WAY CLASSIFICATION ON DATA SET WITH NATURAL DISTRIBUTION; 5-FOLD CROSS-VALIDATION AVERAGES OF PRECISION, RECALL AND F-MEASURE FOR GENERAL OPINION CLASS

Classifier			P	R	F
Binary	GSS	SVM	0.000	0.000	0.000
		BoosTexter	0.000	0.000	0.000
	Unigram	SVM	0.200	0.032	0.055
		BoosTexter	0.416	0.222	0.288
	Unigram + GSS	SVM	0.000	0.000	0.000
		BoosTexter	0.436	0.238	**0.307**
4-way	GSS	BoosTexter	0.000	0.000	0.000
	Unigram	BoosTexter	0.447	0.251	0.320
	Unigram + GSS	BoosTexter	0.463	0.273	**0.341**

Table IV
GENERAL OPINION DETECTION TASK: BINARY AND 4-WAY CLASSIFICATION ON DATA SET WITH BALANCED DISTRIBUTION (RANDOM UNDER-SAMPLING); 5-FOLD CROSS-VALIDATION AVERAGES OF PRECISION, RECALL AND F-MEASURE FOR GENERAL OPINION CLASS

Classifier			P	R	F
Binary	GSS	SVM	0.575	0.508	0.538
		BoosTexter	0.714	0.572	**0.631**
	Unigram	SVM	0.636	0.720	0.671
		BoosTexter	0.652	0.736	**0.691**
	Unigram + GSS	SVM	0.620	0.784	0.689
		BoosTexter	0.686	0.728	**0.703**
4-way	GSS	SVM-MV	0.496	0.408	**0.444**
		BoosTexter	0.429	0.448	0.422
	Unigram	SVM-MV	0.756	0.844	**0.796**
		BoosTexter	0.541	0.552	0.544
	Unigram + GSS	SVM-MV	0.729	0.808	**0.765**
		BoosTexter	0.571	0.596	0.581

Table V
GENERAL OPINION DETECTION TASK: BINARY AND 4-WAY CLASSIFICATION RESULTS ON THE HELD-OUT TEST SET WITH NATURAL DISTRIBUTION; AVERAGE PRECISION, RECALL AND F-MEASURE OF THE MODELS TRAINED ON THE DATA SET WITH NATURAL DISTRIBUTION.

Classifier			P	R	F
Binary	GSS	SVM	0.000	0.000	0.000
		BoosTexter	0.000	0.000	0.000
	Unigram	SVM	0.000	0.000	0.000
		BoosTexter	0.313	0.175	0.224
	Unigram + GSS	SVM	0.000	0.000	0.000
		BoosTexter	0.358	0.175	**0.233**
4-way	GSS	BoosTexter	0.000	0.000	0.000
	Unigram	BoosTexter	0.421	0.256	**0.317**
	Unigram + GSS	BoosTexter	0.375	0.273	0.309

Table VI
GENERAL OPINION DETECTION TASK: BINARY AND 4-WAY CLASSIFICATION RESULTS ON THE HELD-OUT TEST SET WITH NATURAL DISTRIBUTION; AVERAGE PRECISION, RECALL, AND F-MEASURE OF THE MODELS TRAINED ON THE DATA SET WITH THE BALANCED DISTRIBUTION.

Classifier			P	R	F
Binary	GSS	SVM	0.577	0.260	**0.355**
		BoosTexter	0.116	0.600	0.194
	Unigram	SVM	0.124	0.750	0.212
		BoosTexter	0.098	0.590	0.168
	Unigram + GSS	SVM	0.107	0.750	0.187
		BoosTexter	0.102	0.605	0.175
4-way	GSS	SVM-MV	0.072	0.350	0.114
		BoosTexter	0.117	0.436	0.182
	Unigram	SVM-MV	0.097	0.760	0.171
		BoosTexter	0.149	0.457	0.225
	Unigram + GSS	SVM-MV	0.087	0.725	0.154
		BoosTexter	0.157	0.471	0.235

of GSS on the F-measure of the General Opinion category in both – binary and 4-way – settings, for SVM we observe this effect only in the binary setting (recall that 4-way is SVM with Majority Voting). The differences in the performances of BoosTexter are not significant on both the 2-way and the 4-way classification (Paired Two-Tail T-Test: $t(4) = -0.96$ and $t(4) = -0.95$ for 2 and 4-way respectively, both $p > 0.05$). However, for SVM on 2-way classification, GSS significantly improves the performance when used together with unigrams (Paired Two-Tail T-Test: $t(4) = -2.86, p < 0.05$). On the other hand, in the 4-way setting, SVM Majority Voting (SVM-MV), we observe the reverse effect, unigram only classifier performs better. Even though the performance difference between the unigram-only model and the model trained on both features is larger in 4-way classification setting, this difference is not significant (Paired Two-Tail T-Test: $t(4) = 1.73, p > 0.05$).

Unlike the general opinion detection with the natural distribution of the categories, on the balanced data set, the binary classification yields better results than the multi-class one in all the cases, but SVM-MV models utilizing unigrams, i.e. unigram only and unigram + GSS models.

Improved performance of the classifiers trained using GSS feature is expected due to the randomly chosen set of non-general opinion documents. Recall from Figure 1 that general opinion documents show bi-modality, whereas the other categories and the data set as a whole do not. Assuming that the selected samples are representative of their respective categories, i.e. follow the same GSS distribution, in the balanced data set, GSS distribution over all data is expected to show some bi-modality; thus, to have an effect on the classification. In a good sampling, the increase in the performance should be proportional to the ratio between the general opinion and the rest of documents; thus, it is greater in the binary and less in the 4-way classification. This only applies to classifiers "natively" supporting multi-class training (like BoosTexter); in its training phase a classifier learns the data with the distribution of GSS over 1000 documents. The case of Majority Voting is different, it is an ensemble of 3 binary classifiers, each trained on 500 documents; thus, for SVM, in both 2 and 4 way settings a classifier learns similar GSS distributions (it is never exposed to 1000 document GSS distribution).

Tables V and VI present the results of the evaluation on the unseen Test Set, with natural distribution. When one compares results in Table V with the results in Table III, one can observe that training with natural distribution of classes is predictive of the results. Even though the performances

Table VII
OPINION POLARITY CLASSIFICATION TASK: AVERAGE ACCURACIES ON GENERAL OPINION-ONLY DATA SET (5-FOLD CROSS-VALIDATION). NATURAL DISTRIBUTION TRAINING (UPPER HALF); BALANCED DISTRIBUTION TRAINING (LOWER HALF, RANGE 0-10); AND BALANCED DISTRIBUTION TRAINING WITH REMOVAL OF AMBIGUOUS RATING DOCUMENTS (LOWER HALF, RANGE 1-4,7-10)

Data Set	Classifier			
GSS Range		GSS	Unigram	Comb.
Natural Distribution				
0-10	SVM	0.835	0.703	0.842
0-10	BoosTexter	0.813	0.758	0.819
Balanced Distribution				
0-10	SVM	0.791	0.644	0.804
0-10	BoosTexter	0.778	0.713	0.804
1-4,7-10	SVM	0.875	0.588	0.875
1-4,7-10	BoosTexter	0.875	0.688	0.838

Table VIII
OPINION POLARITY CLASSIFICATION TASK: AVERAGE ACCURACIES ON THE HELD-OUT GENERAL OPINION-ONLY TEST SET. NATURAL DISTRIBUTION TRAINING (UPPER HALF); BALANCED DISTRIBUTION TRAINING (LOWER HALF, RANGE 0-10); AND BALANCED DISTRIBUTION TRAINING WITH REMOVAL OF AMBIGUOUS RATING DOCUMENTS (LOWER HALF, RANGE 1-4,7-10)

Data Set	Classifier			
GSS Range		GSS	Unigram	Comb.
Natural Distribution				
0-10	SVM	0.805	0.525	0.795
0-10	BoosTexter	0.745	0.595	0.745
Balanced Distribution				
0-10	SVM	0.805	0.565	0.820
0-10	BoosTexter	0.780	0.590	0.775
1-4,7-10	SVM	0.825	0.530	**0.830**
1-4,7-10	BoosTexter	0.825	0.510	0.765

on the test set are generally lower, the performance of the BoosTexter unigram only model during training is very close to the obtained results (F-measures 0.320 *vs* 0.317). However, on the test set, we fail to observe the positive effect of GSS.

Comparing the results of training on balanced data and testing on natural distribution (Tables IV and VI), the first observation is that the high performance of classifiers disappears.

By comparing Tables V and VI, one can assess the effects of balancing the data set for training: (1) Altering the GSS distribution makes it possible for GSS only classifiers to yield non-zero results. (2) Classifiers learn a high-recall / low precision models using unigram features. (3) Reducing the number of documents results in the deterioration of unigram only model performances.

Having received these results we can conclude that training on a data with natural distribution of labels is the most predictive of the performance on the unseen data with an unknown distribution (The distribution of labels in future surveys might be different; however, it will be closer to the natural distribution, rather than the balanced one.) Moreover, multi-class training results are the closest to the test set performance.

C. Polarity Classification Task

In this section we present results of experiments on Polarity Classification. Polarity Classification is a binary classification task, where the data consists only of documents annotated as a general opinion by the expert. The general opinion class of documents is not very numerous compared to the whole training set: close to 8% (316 documents): 62.3% (197) of which are labeled as negative and 37.3% (118) as positive, and 1 document is labeled as both; thus, it is removed for the polarity classification.

Each of the described experiments is performed on different subsets of the general opinion data set. Besides training with natural distribution, we apply different random undersampling settings to obtain balanced general opinion-only data set. The selection criterion is the user rating scale. The first setting is to sample from documents that belong to the whole 0-10 range, i.e. all documents. The second setting is to sample from documents that belong to either 1-4 or 7-10 user rating ranges, that is we remove documents that belong to the ambiguous "middle" part. Removing the "middle" part is expected to improve the performance. As a result, the data set obtained using the first setting contains 115 documents of each polarity; removing the "middle" part yields a smaller data set – 80 document of each polarity.

Since in the binary classification, where both classes are of interest, accuracy is equivalent to micro-averaged F-measure, the average accuracies of 5-fold cross-validation results are reported, rather than micro-averaged precision, recall, and F-measure.

The Table VII (the first line of results) displays the results on 5-fold cross-validation classification on the 2-way polarity classification on 315 document general opinion set. GSS only classifier performs better than the unigram based classifier, which is expected given the distribution of GSS on Figure 1 and the small size of the general opinion data set. Both for BoosTexter and SVM models, the combination of unigram and GSS features outperforms the models using GSS or unigram features alone. However, the observed improvement is not significant with respect to GSS only classification (Paired Two-Tail T-Test, BoosTexter: $t(4) = -0.22, p > 0.05$, SVM: $t(4) = -0.33, p > 0.05$).

The results in Table VII indicate that balancing the data set has a negative effect on unigram only classifiers: the more data is removed, the worse the performance. As expected, GSS only classifier performances increase only after removing the "ambiguous" middle part: just balancing the data set decreases the performance. In parallel to this, combining unigram and GSS features improves performance over GSS only classification only in the case of the balanced data set that still contains "ambiguous" middle range documents. This effect disappears when the middle range documents are removed.

Our results on opinion polarity classification indicate that

it is indeed possible to assign polarity labels based on the user ratings only; however, the opinion detection step is required. Moreover, even if the "ambiguous" middle part is removed from the data set, the noise is still to be expected.

Similar to the polarity classification in training phase, we extract all the general opinion documents from the held-out test set to form a separate general opinion-only test set (40 documents). In this test set 12 (30%) of documents are labeled as negative and the remaining 28 (70%) as positive.

When one compares results in Tables VII and VIII, one can observe that for both – the natural distribution and balanced distribution training – GSS only classifier group is the most predictive. Moreover, removing the "ambiguous" middle range documents improves the performance.

In balanced training, SVM benefits from unigram features as well: the combined performance is generally higher; however, the difference is not significant. From the results on Polarity Classification, we conclude that GSS only classifier trained on balanced data with removed "ambiguous" range can predict document polarity sufficiently well. Thus, unsupervised classification of documents into positive and negative polarities within the general opinion group using thresholding is reasonable.

VI. Conclusion

We presented several experiments for detecting general opinion documents by complementing unigram features from free-text comments with the user rating feature from the questionnaire. However, we observed that user ratings are not useful for detection of general opinions. Thus, we conclude that general opinion detection works better from the textual features only. Another observation is that balancing the data set has the negative effect when evaluated on a test set with natural distribution. Thus, it should be avoided unless there is a way to balance unseen data in an unsupervised manner.

For the opinion polarity classification we observed that user ratings are very effective , which validates the approach used in the literature to use them for inferring polarity; however, a document must first be identified as a general opinion.

Acknowledgment

We would like to thank Paolo Menardi and Paola Malesardi for their technical insights and Marco La Manna and Pietro Parente for his support. This work was partly funded by a Telecom Italia research grant.

References

[1] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the ACL*, 2005, pp. 115–124.

[2] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the trec-2006 blog track," in *Proceedings of TREC*, 2006.

[3] J. Zhu, H. Wang, M. Zhu, B. K. Tsou, and M. Ma, "Aspect-based opinion polling from customer reviews," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 37 –49, 2011.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of EMNLP*, Philadelphia, PA, 2002, pp. 79–86.

[5] T. Bhuiyan, Y. Xu, and A. Josang, "State-of-the-art review on opinion mining from online customers' feedback," in *Proceedings of the 9th Asia-Pacific Complex Systems Conference*, 2009.

[6] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of ACL*, 2007, pp. 976–983.

[7] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," in *Proceedings of EACL*, 2006.

[8] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of EMNLP*, 2003.

[9] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of COLING*, 2004, pp. 841–847.

[10] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of ACL*, 2007.

[11] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the ACL*, 2004.

[12] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? finding strong and weak opinion clauses," in *Proceedings of the 19th national conference on Artificial intelligence*, 2004, pp. 761–767.

[13] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings 19th National Conference on Artificial Intelligence*, 2004.

[14] H. Binali, V. Potdar, and C. Wu, "A state of the art opinion mining and its application domains," in *Proceedings of IEEE International Conference on Industrial Technology*, 2009.

[15] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[16] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT-Press, 1999.

[17] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[18] R. Yan, Y. Liu, R. Jin, and A. Hauptmann, "On predicting rare classes with svm ensembles in scene classification," in *Proceedings of ICASSP*, 2003.