Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference

Genre-Based Stages Classification for Polarity Analysis

John Roberto, Maria Salamó, and **M. Antònia Martí**
University of Barcelona
Gran Via 585, 08007 Barcelona, Spain
{roberto.john,maria.salamo,amarti}ub.edu

Abstract

Polarity detection of Online Reviews is one of the most popular tasks related to Opinion Mining. Given that most state-of-the-art solutions ignore the structural aspects of a review, we present an approach to polarity detection that, first, distinguishes stages in the genre of hotel reviews and, subsequently, evaluates the usefulness of each type of stage in the determination of the polarity of the entire review. Our experiments show that our proposal provides good accuracy rates in identifying the overall polarity of a review by using a very small proportion of the text.

Introduction

Research on Opinion Mining focuses on classifying the polarity (positive, negative) of opinionated texts such as reviews from the perspective of textual evidence. Most of the work in the field has been intensively applied on the English language but Spanish is required. In this article, we describe a genre-based method for the polarity analysis of Spanish reviews. The genre of texts has been considered by many authors to be a relevant factor in the detection of affect and emotion (Pajupuu, Kerge, and Altrov 2012; Li et al. 2012). In a broad sense, Swales (1990) considers that each genre is characterized by a "schematic structure" composed of different types of stages, each one with a goal-oriented function. In functional theories of discourse, stages are typical sequences of sentences (or paragraphs) that characterize the information structure (local and global) of different text genres. We propose to split reviews into different types of functional stages and subsequently select only the most relevant ones in order to obtain their overall polarity.

Stages in Hotel Reviews

Customer reviews in general are not overly complex in structure. According to Ricci and Wietsma (2006), a review can be defined as a subjective piece of text describing user experiences, product knowledge and opinions –together with a final product rating. Each of these types of information plays a distinct function in achieving the overall purpose of the customer review and, consequently, they can be represented by a set of functional types of stages (Martin 1993; Swales 1990).

In our view, the first type of information, experiences, is related to general information about the user, for example: *"Having got married last week my new husband and I went for a few days to the ..."*. This information is not restricted to any domain. In contrast, the second type of information, product descriptions (subjective and objective descriptions), is related to the customer experience with a specific product, for example: *"the image quality and general shooting performance are top-notch"*. In accordance with marketing research theories, this information is largely based on cognitive learning and coupled with credible experience with many offerings and brands within a product category. Finally, the third type of information, opinions, is associated with user self-awareness and introspection, for example: *"I would buy this phone again"*. We argue that these three types of stages are present in a standard hotel review, and that they do not contribute in the same way to the expression of polarity.

Data

With the objective of evaluating the usefulness of each type of stage to determine the polarity of the reviews, we collected 150 reviews from Tripadvisor.com. This corpus was hand-labeled with stage types and enriched with a variety of linguistic data (e.g. PoS, domain and stage terms), which were provided by automatic or semi-automatic annotation mechanisms.

Once the corpus was annotated, we analyzed the distribution of the stages of the reviews. Figure 1 shows the breakdown of high-level stages for the 150 reviews. Figure 1A groups the reviews with the same stage patterns. The majority of reviews are grouped around the sequence *narrative* > *descriptive* > *introspective* (N-D-I), which can be referred to as the prototypical or canonical pattern in hotel reviews. In Figure 1B we have grouped the reviews according to the number of stages that they contain. We validate the reliability of manual annotation by applying machine learning techniques (see section below).

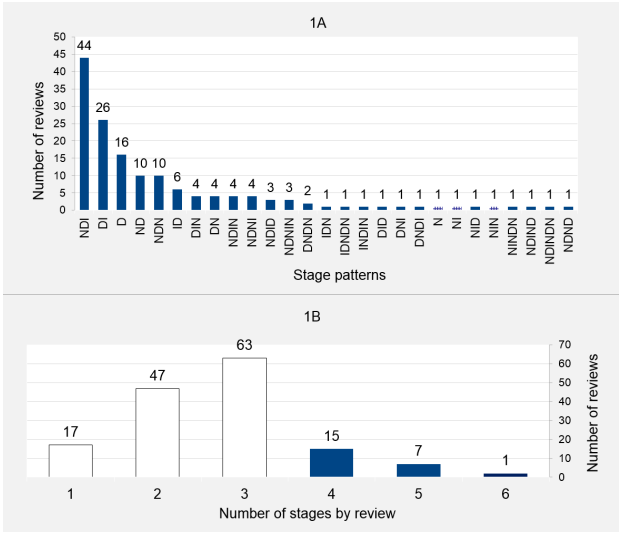

Figure 1: Reviews grouped by number and distribution of stages.

Experiments and Results

We evaluated our hypothesis with two experiments: stage categorization and polarity analysis. The purpose of the first experiment was to detect and characterize the three types of stages: narrative, descriptive and introspective. The second experiment assessed the usefulness of the different type of stages to improve opinion polarity classification.

Stage Categorization

In this experiment, we applied a machine learning approach to the automatic detection and characterization of stage types in hotel reviews using a pre-established set of features. Stages were first transformed into a representation suitable for the application of classification algorithms. In our study, we used a set of features, F, that characterize the set of stages, S. We constructed a matrix M setting out the stages as rows and the features as columns: $M = \{S \text{ x } F\}$. The maximum number of stages in the document collection is 450 (150 reviews x 3 possible types of stages per review), therefore $S = \{s_1 ... s_{\leq 450}\}$. It should be noted that for each review we put together stages with the same label. For example, in the sequence *descriptive* > *narrative* > *descriptive* we account for 2 stages (types).

We propose a set of 14 features which best distinguish between types of stages: $F = \{f_1 ... f_{14}\}$. In the following, we list these features and we discuss the theoretical background that leads us to propose each feature:

[f_1] **Word count per stage:** we hypothesized that there is a significant difference in the number of words in each type of stage: description and narration stages tend to be longer than the introspective stage.

[f_2] **Tense and mood:** we hypothesized that conditional and subjunctive forms of verbs, together with the future tense, are characteristic of introspective stages (e.g. *no se lo recomendaría a nadie* "I would not recommend it to anybody").

[f_3 to f_5] **Grammatical person (1st, 2nd and 3rd):** we hypothesized that narratives are written basically in first person (e.g. *pasamos dos días agradables* "we spent two nice days"), descriptions in third person (e.g. *el hotel está en una ubicación muy apropiada* "the hotel is at a very convenient location") and introspections in first person (e.g. *no me plantearía alojarme en ningún otro sitio* "I would not consider staying anywhere else") or second person (e.g. *tendrías que considerar otras opciones* "you should certainly consider other options").

[f_6] **Domain-specific terms:** we hypothesized that domain-specific terms occur more often in descriptive stages (e.g. *personal* "staff", *registrarse* "check-in").

[f_7 and f_8] **Stage-specific terms (single words and trigrams):** we hypothesized that there are single words and trigrams that characterize each type of stage in accordance with the type of information that they contain: narrative (e.g. *casados* "married", *por tres dias* "for three days"), descriptive (e.g. *precioso* "gorgeous", *muy cerca de* "very close to") and introspective (e.g. *desear* "wish", *mi próximo viaje* "my next trip").

[f_9 to f_{12}] **Lexical aspect of verbs (accomplishments, achievements, states and activities)**[1]**:** we hypothesized that typical activity verbs (e.g. *caminar* "walk") are more common in narrative stages, while typical stative verbs (e.g. *tener* "have") are more frequent in descriptive stages.

[f_{13}] **Verb frequency:** we hypothesized that verb frequency in narrative stages is higher than in descriptive stages.

[f_{14}] **Summing-up discourse markers:** we hypothesized that discourse markers expressing conclusion or summary are typical of introspective stages (e.g. *en resumen* "in short").

The resulting matrix of features, M, contains m rows and f columns, where each row, m_i corresponds to an example of stage type and each column f_j is one of the fourteen features previously defined. Accordingly, for simplifying notation, we use m_{ij} to refer to the value of the i^{th} example for the j^{th} feature. We performed a linear transformation on the original data to scale the value of all features in the range [0..1]. The normalization is calculated by the formula $Norm(m_{ij}) = m_{ij} - min(f_j)/max(f_j) - min(f_j)$, where m_{ij} is the current value of the ith example for jth feature, $min(f_j)$ is the minimum value for feature f_j among all examples, and $max(f_j)$ is the maximum value for feature f_j among all examples.

We conducted our experiments using the Weka framework. We experimented with four mainstream classification algorithms, two feature selection methods, and one search method –the complete list is available at the bottom of Table 1. The dataset was split into training and test sets using a stratified 10-fold cross-validation for the four possible stage configurations: N∩D, N∩I, D∩I, and N∩D∩I.

The results are summarized in Table 1. The first column shows the four different configurations of the target con-

[1] Polysemous verbs were not disambiguated.

cepts: narrative versus descriptive versus introspective, narrative versus descriptive and so on. Column two contains the prediction accuracy score ($Accuracy$), which is simply the total number of stages correctly classified, obtained for each of the four stage configurations. The third column contains the three most relevant features picked up by attribute selection methods, which were ranked across all folds in each stage configuration. Finally, the last column contains the feature identifier ($\#f$) in accordance with the list stated above. An asterisk (*) means that both features are equally relevant. The last row in the table shows the overall average ($Average$) for all configurations.

Stages	Accuracy	Features	$\#f$
		Stage-specific terms (words)	f_7
N∩D∩I	81.4 %	Activity verbs*	f_{12}
		Domain-specific terms*	f_6
		Stage-specific terms (words)	f_7
N∩D	97.4 %	Activity verbs*	f_{12}
		Domain-specific terms*	f_6
		Stage-specific terms (words)*	f_7
N∩I	83.7 %	Verb frequency*	f_{13}
		First person	f_3
		Domain-specific terms	f_6
D∩I	93.1 %	First person	f_3
		Third person	f_5
Average	**88.9 %**		

Classification algorithms: 1. Bayes (BayesNet, DMNBtext) 2. Lazy (IBk, KStar, LWL) 3. Rules (ConjunctiveRule, DTNB, DecisionTable, JRip, OneR, PART) 4. Trees (ADTree, BFTree, J48, J48graft, LMT, NBTree, REPTree, RandomForest)
Selection methods: Information Gain (IG) and χ^2 (CHI-SQUARE))
Search method: Ranker

Table 1: Stage categorization performance of Spanish reviews

The results indicate that the highest accuracy (97.4%) on average is obtained when the classes to be learned are narrative and descriptive (N∩D configuration). This is because narrations and descriptions are very different from one another in content and function. We achieved a good degree of accuracy (93.1%) when the two classes to be learned were descriptive and introspective stages (D∩I configuration). This accuracy is not as good as the previous configuration because the introspective stage sometimes "summarizes" the descriptive stage and, in consequence, they both share some features such as the vocabulary (note that the stage-specific terms feature is not relevant here). For the classification of the narrative and introspective stages we obtained an accuracy of 83.7%, this value is lower than N∩D but higher than the N∩D∩I configuration. In the latter case, we have shown that it is possible to reach accuracy rates as high as 81%.

Let us now consider the relevance of features. *Stage-specific terms* is the most relevant feature in three of the four configuration or stage classification tasks. **D**∩I configuration does not use *stage-specific terms* because, as we noted above, the descriptive and introspective stages share some vocabulary. *Domain-specific terms* characterize descriptive stages: this feature appears in all configurations involving descriptions (N∩**D**∩I, N∩**D**, **D**∩I). This is not surprising given that –in accordance with our hypothesis– a descriptive stage contains information about the domain. *Verbs* have been shown to be useful in distinguishing descriptions from narratives: *activity verbs* in **N**∩D∩I and **N**∩D configurations, and *verbs frequency* in **N**∩I. Verbs are relevant here because narrative stages report general information about customers in the form of short stories. The introspective stage (N∩**I** and D∩**I** configurations) is characterized by the "first person" since it is usually concerned with customer self-awareness.

The results obtained confirm that it is possible to achieve good performance in identifying the narrative, descriptive and introspective stages in hotel reviews. The next step consists of the use of those stages to perform polarity analysis.

Polarity Analysis

The second experiment consists of determining the usefulness of the different type of stages for the classification of positive and negative reviews (polarity analysis).

For polarity analysis, texts were modeled as a matrix with reviews as rows and words as columns, *M = R* x *W*. In this experiment we worked with 120 reviews, *R* = $\{r_1$ ... $r_{120}\}$, since neutral reviews (those with a rating of 3) were not taken into consideration. Each $r_n \in$ R is represented by a finite set of words or bag-of-words (BoW). In particular, there were three different instances of BoW in accordance with the type of Stage (S) to be represented: descriptive *BoW*$_d$ = $\{w_{d1}$ ... $w_{dt}\}$, narrative *BoW*$_n$ = $\{w_{n1}$ ... $w_{nt}\}$ and introspective *BoW*$_i$ = $\{w_{i1}$ ... $w_{it}\}$. Additionally, another BoW_a was used in order to contain all the BoW representations, *BoW*$_a$ = $\{w_{a1}$ ... $w_{at'}\}$ where *t'* = $|BoW_d|$ + $|BoW_n|$ + $|BoW_i|$. The latter representation corresponds to the *benchmark scenario* because it uses all the available data.

Additionally, three term weighting schemes were used, namely, $binary$, tf and $tf - idf$. The binary scheme only considers whether a term t appears in the review representation. The raw term frequency (tf) is defined as the number of times a term t appears in the review representation. Inverse document frequency $tf - idf$ is based on counting the number of reviews in the collection being searched, which contain (or are indexed by) the term t.

In this experiment we tested 37 Weka machine learning algorithms grouped into six mainstream categories –the list of implemented algorithms appears at the bottom of Table 2– and, again, we evaluated each algorithm using a stratified ten-fold-cross validation process. In total, 4440 models were built in our case study: 10 folds * 37 algorithms * 3 weighting schemes (TWS) * 4 review representation (narrative, descriptive, introspective and benchmark). Table 2 presents the average prediction accuracy, the vocabulary size (the number of words used to build each review representation) and the performance ratio for the four review representations: the narrative $M_n = (R$ x $BoW_n)$, the descriptive $M_d = (R$ x $BoW_d)$, the introspective $M_i = (R$ x $BoW_i)$, and the benchmark $M_a = (R$ x $BoW_a)$. Prediction accuracy (PA) is displayed in relation to vocabulary size. Therefore, in our

analysis the performance ratio (PR) is obtained by dividing the (best) accuracy by the vocabulary size. For example, the best accuracy for M_i representation (70.96%) divided by its vocabulary size (10.81%) gives a PR of 6.56.

Both prediction accuracy (PA) and performance ratio (PR[2]) are two complementary benchmarks to measure the quality of our classifiers. We give special relevance to PR because this metric gives us useful information about the vocabulary data reduction induced by the use of the stage types.

	Review representation (PA)			
Term weighting schemes	M_n	M_d	M_i	benchmark M_a
binary	64.39%	63.53%	**70.96%**	67.81%
tf	66.08%	65.09%	70.63%	74.18%
tf-idf	63.44%	68.91%	68.51%	73.08%
Vocabulary size	23.98%	65.21%	10.81%	100%
Performance Ratio	2.76	1.06	**6.56**	0.74

Classification algorithms: 1. Bayes (BayesNet, BayesianLogisticRegression, ComplementNaiveBayes, DMNBtext, NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable, NaiveBayesUpdateable), 2. Lazy (IB1, IBk, KStar, LWL), 3. Misc (HyperPipes, VFI), 4. Rules (ConjunctiveRule, DTNB, DecisionTable, JRip, NNge, OneR, PART, Ridor, ZeroR), 5. Trees (ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, NBTree, REPTree, RandomForest, RandomTree, SimpleCart, lmt.LogisticBase) and 6. functions.SMO

Table 2: Item polarity performance of Spanish reviews

Regarding the absolute categorization accuracy of data (PA), the M_i review representation has the best accuracy (70.9%) in comparison with the other two representations based on review segmentation, M_d (68.9%) and M_n (66%). This PA is only 3.2% below the best accuracy obtained with the benchmark M_a which uses all the available data (the full review). In contrast, the difference with respect to M_a for M_d and M_n is higher: 5.2% and 8.1%, respectively. On average the M_i review representation does not reduce accuracy as much as the M_d and M_n representations do.

The major significant difference between review representations was found by bringing together PA and vocabulary size in the PR. The reduction in vocabulary size for the M_i review representation is 89.19% because we only used 10.81% of the data. The reduction in vocabulary size from the benchmark (M_a) to M_n is also large (76%). By contrast, for M_d representation, more than 65% of the words are found in M_a, which means a reduction of only 34% in relation to M_a. The best PR is for the M_i representation, with 6.56%. This ratio is significantly higher than the same ratio in M_a, the benchmark representation. M_n is also higher than M_a but, as we commented before, its accuracy is not good enough. Performance ratio for M_d is not meaningful. These findings clearly indicate that there is a good balance between accuracy and vocabulary size.

Additionally, we want to emphasize the dependence observed between term weighting schemes, review representations and vocabulary size in Table 2: the descriptive stages, which have the largest vocabulary size, are better represented by the $tf - idf$ scheme; the narrative stages, which have a medium vocabulary size, are better represented by the tf scheme; and the introspective stages, which have the smallest vocabulary size, are better represented by the $binary$ scheme. It is important to state that the $binary$ is the most straightforward scheme and this property is very important in order to improve introspective stage representation performance (M_i).

These findings allow us to conclude that it is possible to obtain a good performance in polarity prediction by using only the introspective stage in conjunction with $binary$ data representation: we obtained an accuracy close to 71% with only 10.8% of the data. This accuracy is very similar to that observed for M_a (benchmark), which, in contrast, uses a very high-dimensional and sparse data matrix.

[2]PR is related to *vocabulary size* and *feature vector length* reduction methods (Arafat et al. 2014).

Conclusions

The main objective of this paper is to detect and assess the usefulness of different types of stages to determine the overall polarity of hotel reviews. In the first experiment, we showed that it is possible to obtain an accuracy of 88.9% in the automatic detection of stages by using a simple set of features. In a second experiment, we demonstrated that, in a basic model, it is also possible to report a good degree of accuracy in polarity categorization by using only the introspective stages (70.9% of accuracy with 10.8% of the data) instead of the whole review (74.1% of accuracy with 100% of the data). As a general conclusion, we like to remark that selecting the most efficient part of the text is fundamental for the optimization of the polarity analysis of reviews.

Acknowledgments

This work was supported by projects SGR-2014-623, TIN2012-38603-C02-02 and TIN2012-38584-C06-01, as well as a FI grant (2010FI_B 00521).

References

Arafat, H.; Elawady, R.; Barakat, S.; and Elrashidy, N. 2014. Different feature selection for sentiment classification. *IJISIS* 3(1):137–150.

Li, H.; Chen, Y.; Ji, H.; Muresan, S.; and Zheng, D. 2012. Combining social cognitive theories with linguistic features for multi-genre sentiment analysis. In *PACLIC26*, 127–136.

Martin, J. 1993. *The Powers of Literacy: a genre approach to teaching writing*. London: Falmer Press.

Pajupuu, H.; Kerge, K.; and Altrov, R. 2012. Lexicon-based detection of emotion in different types of texts. In *Estonian Papers in Applied Linguistics*, 171–184.

Ricci, and Wietsma. 2006. Product reviews in travel decision making. In *Information and Communication Technologies in Tourism 2006*, 296–307.

Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.