Liars and Saviors
in a Sentiment Annotated Corpus of Comments to Political debates

Paula Carvalho
University of Lisbon
Faculty of Sciences, LASIGE
Lisbon, Portugal
pcc@di.fc.ul.pt

Luís Sarmento
Labs Sapo UP & University of Porto
Faculty of Engineering, LIACC
Porto, Portugal
las@co.sapo.pt

Jorge Teixeira
Labs Sapo UP & University of Porto
Faculty of Engineering, LIACC
Porto, Portugal
jft@fe.up.pt

Mário J. Silva
University of Lisbon
Faculty of Sciences, LASIGE
Lisbon, Portugal
mjs@di.fc.ul.pt

Abstract

We investigate the expression of opinions about human entities in user-generated content (UGC). A set of 2,800 online news comments (8,000 sentences) was manually annotated, following a rich annotation scheme designed for this purpose. We conclude that the challenge in performing opinion mining in such type of content is correctly identifying the positive opinions, because (i) they are much less frequent than negative opinions and (ii) they are particularly exposed to verbal irony. We also show that the recognition of human targets poses additional challenges on mining opinions from UGC, since they are frequently mentioned by pronouns, definite descriptions and nicknames.

1   Introduction

Most of the existing approaches to opinion mining propose algorithms that are independent of the text genre, the topic and the target involved. However, practice shows that the opinion mining challenges are substantially different depending on these factors, whose interaction has not been exhaustively studied so far.

This study focuses on identifying the most relevant challenges in mining opinions targeting media personalities, namely politicians, in comments posted by users to online news articles. We are interested in answering open research questions related to the expression of opinions about human entities in UGC.

It has been suggested that the target identification is probably the easiest step in mining opinions on products using product reviews (Liu, 2010). But, is this also true for human targets namely for media personalities like politicians? How are these entities mentioned in UGC? What are the most productive forms of mention? Is it a standard name, a nickname, a pronoun, a definite description? Additionally, it was demonstrated that irony may influence the correct detection of positive opinions about human entities (Carvalho et al., 2009); however, we do not know the prevalence of this phenomenon in UGC. Is it possible to establish any type of correlation between the use of irony and negative opinions? Finally, approaches to opinion mining have implicitly assumed that the problem at stake is a balanced classification problem, based on the general assumption that positive and negative opinions are relatively well distributed in

564

texts. But, should we expect to find a balanced number of negative and positive opinions in comments targeting human entities, or should we be prepared for dealing with very unbalanced data?

To answer these questions, we analyzed a collection of comments posted by the readers of an online newspaper to a series of 10 news articles, each covering a televised face-to-face debate between the Portuguese leaders of five political parties. Having in mind the previously outlined questions, we designed an original rich annotation scheme to label opinionated sentences targeting human entities in this corpus, named *SentiCorpus-PT*. Inspection of the corpus annotations supports the annotation scheme proposed and helps to identify directions for future work in this research area.

2   Related Work

MPQA is an example of a manually annotated sentiment corpus (Wiebe et al., 2005; Wilson et al., 2005). It contains about 10,000 sentences collected from world press articles, whose private states were manually annotated. The annotation was performed at word and phrase level, and the sentiment expressions identified in the corpus were associated to the source of the private-state, the target involved and other sentiment properties, like intensity and type of attitude. MPQA is an important resource for sentiment analysis in English, but it does not reflect the semantics of specific text genres or domains.

Pang et al. (2002) propose a methodology for automatically constructing a domain-specific corpus, to be used in the automatic classification of movie reviews. The authors selected a collection of movie reviews where user ratings were explicitly expressed (e.g. "4 stars"), and automatically converted them into positive, negative or neutral polarities. This approach simplifies the creation of a sentiment corpus, but it requires that each opinionated text is associated to a numeric rating, which does not exist for most of opinionated texts available on the web. In addition, the corpus annotation is performed at document-level, which is inadequate when dealing with more complex types of text, such as news and comments to news, where a multiplicity of sentiments for a variety of topics and corresponding targets are potentially involved (Riloff and Wiebe., 2003; Sarmento et al., 2009).

Alternative approaches to automatic and manual construction of sentiment corpora have been proposed. For example, Kim and Hovy (2007) collected web users' messages posted on an election prediction website (www.electionprediction.org) to automatically build a gold standard corpus. The authors focus on capturing lexical patterns that users frequently apply when expressing their predictive opinions about coming elections. Sarmento et al. (2009) design a set of manually crafted rules, supported by a large sentiment lexicon, to speed up the compilation and classification of opinionated sentences about political entities in comments to news. This method achieved relatively high precision in collecting negative opinions; however, it was less successful in collecting positive opinions.

3   The Corpus

For creating *SentiCorpus-PT* we compiled a collection of comments posted by the readers of the Portuguese newspaper Público to a series of 10 news articles covering the TV debates on the 2009 election of the Portuguese Parliament. These took place between the 2nd and the 12th of September, 2009, and involved the candidates from the largest Portuguese parties. The whole collection is composed by 2,795 posts (approx. 8,000 sentences), which are linked to the respective news articles.

This collection is interesting for several reasons. The opinion targets are mostly confined to a predictable set of human entities, i.e. the political actors involved in each debate. Additionally, the format adopted in the debates indirectly encouraged users to focus their comments on two specific candidates at a time, persuading them to confront their standings. This is particularly interesting for studying both direct and indirect comparisons between two or more competing human targets (Ganapathibhotla and Liu, 2008).

Our annotation scheme stands on the following assumptions: (i) the sentence is the unit of analysis, whose interpretation may require the analysis of the entire comment; (ii) each sentence may convey different opinions; (iii) each opinion may have different targets; (iv) the targets, which can be omitted in text, correspond to human entities; (v) the entity mentions are classifiable into syntactic-semantic categories; (vi) the opinionated sentences may be characterized according to their polarity

and intensity; (vii) each opinionated sentence may have a literal or ironic interpretation.

Opinion Target: An opinionated sentence may concern different opinion targets. Typically, targets correspond to the politicians participating in the televised debates or, alternatively, to other relevant media personalities that should also be identified (e.g. *The Minister of Finance is done!*). There are also cases wherein the opinion is targeting another commentator (*e.g. Mr. Francisco de Amarante, did you watch the same debate I did?!?!?*), and others where expressed opinions do not identify their target (e.g. *The debate did not interest me at all!*). All such cases are classified accordingly.

The annotation also differentiates how human entities are mentioned. We consider the following syntactic-semantic sub-categories: (i) *proper name*, including acronyms (e.g. *José Sócrates, MFL*), which can be preceded by a title or position name (e.g. *Prime-minister José Sócrates; Eng. Sócrates*); (ii) *position name* (e.g. *social-democratic leader*); (iii) *organization* (e.g. *PS party, government*); (iv) *nickname* (e.g. *Pinócrates*); (v) *pronoun* (e.g. *him*); (vi) *definite description*, i.e. a noun phrase that can be interpreted at sentence or comment level, after co-reference resolution (*e.g. the guys at the Ministry of Education*); (vii) *omitted*, when the reference to the entity is omitted in text, a situation that is frequent in null subject languages, like European Portuguese (e.g. *[He] massacred...*).

Opinion Polarity and Intensity: An opinion polarity value, ranging from «-2» (the strongest negative value) to «2» (the strongest positive value), is assigned to each of the previously identified targets. Neutral opinions are classified with «0», and the cases that are ambiguous or difficult to interpret are marked with «?».

Because of its subjectivity, the full range of the intensity scale («-2» vs. «-1»; «1» vs. «2») is reserved for the cases where two or more targets are, directly or indirectly, compared at sentence or comment levels (e.g. *Both performed badly, but Sócrates was clearly worse*). The remaining negative and positive opinions should be classified as «-1» and «1», respectively.

Sentences not clearly conveying sentiment or opinion (usually sentences used for contextualizing or quoting something/someone) are classified as «non-opinionated sentences».

Opinion Literality: Finally, opinions are characterized according to their literality. An opinion can be considered literal, or ironic whenever it conveys a meaning different from the one that derives from the literal interpretation of the text (e.g. *This prime-minister is wonderful! Undoubtedly, all the Portuguese need is more taxes!*).

4 Corpus Analysis

The *SentiCorpus-PT* was partially annotated by an expert, following the guidelines previously described. Concretely, 3,537 sentences, from 736 comments (27% of the collection), were manually labeled with sentiment information. Such comments were randomly selected from the entire collection, taking into consideration that each debate should be proportionally represented in the sentiment annotated corpus.

To measure the reliability of the sentiment annotations, we conducted an inter-annotator agreement trial, with two annotators. This was performed based on the analysis of 207 sentences, randomly selected from the collection. The agreement study was confined to the target identification, polarity assignment and opinion literality, using Krippendorff's Alpha standard metric (Krippendorff, 2004). The highest observed agreement concerns the target identification ($\alpha=0.905$), followed by the polarity assignment ($\alpha=0.874$), and finally the irony labeling ($\alpha=0.844$). According to Krippendorff's interpretation, all these values (> 0.8) confirm the reliability of the annotations.

The results presented in the following sections are based on statistics taken from the 3,537 annotated sentences.

4.1 Polarity distribution

Negative opinions represent 60% of the analyzed sentences. In our collection, only 15% of the sentences have a positive interpretation, and 13% a neutral interpretation. The remaining 12% are non-opinionated sentences (10%) and sentences whose polarity is vague or ambiguous (2%). If one considers only the elementary polar values, it can be observed that the number of negative sentences is about three times higher than the number of positive sentences (68% vs. 17%).

The graphic in Fig. 1 shows the polarity distribution per political debate. With the exception of the debate between Jerónimo de Sousa (C5) and

Paulo Portas (C3), in which the number of positive and negative sentences is relatively balanced, all the remaining debates generated comments with much more negative than positive sentences.
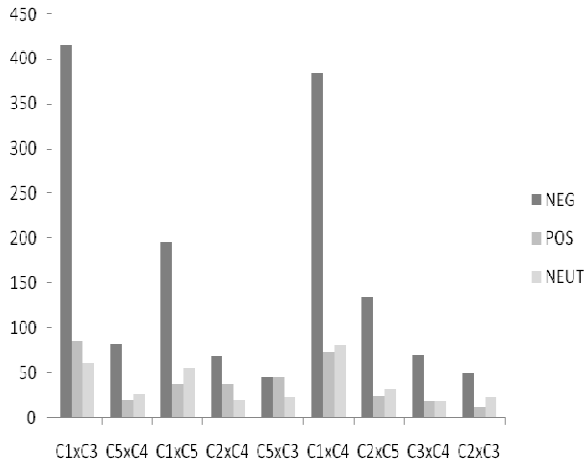

Fig. 1. Polarity distribution per political debate

When focusing on the debate participants, it can be observed that José Sócrates (C1) is the most censured candidate, and Jerónimo de Sousa (C5) the least censured one, as shown in Fig. 2. Curiously, the former was reelected as prime-minister, and the later achieved the lowest percentage of votes in the 2009 parliamentary election.
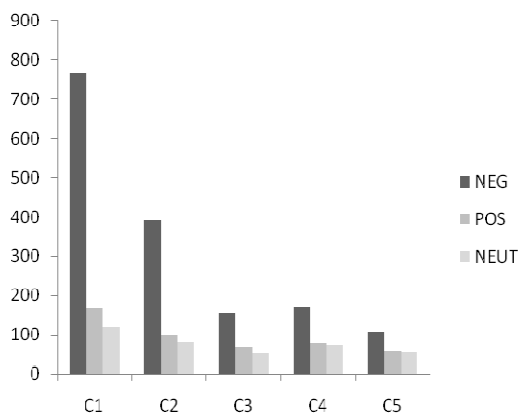

Fig. 2. Polarity distribution per candidate

Also interesting is the information contained in the distributions of positive opinions. We observe that there is a large correlation (The Pearson correlation coefficient is $r = 0.917$) between the number of comments and the number of votes of each candidate (Table 1).

Candidate (C)	#PosCom	#Votes
José Sócrates (C1)	169	2,077,238
M. Ferreira Leite (C2)	100	1,653,665
Paulo Portas (C3)	69	592,778
Francisco Louçã (C4)	79	557,306
Jerónimo de Sousa (C5)	58	446,279

Table 1. Number of positive comments and votes

4.2 Entity mentions

As expected, the most frequent type of mention to candidates is by name, but it only covers 36% of the analyzed cases. Secondly, a proper or common noun denoting an organization is used metonymically for referring its leaders or members (17%). Pronouns and free noun-phrases, which can be lexically reduced (or omitted) in text, represent together 38% of the mentions to candidates. This is a considerable fraction, which cannot be neglected, despite being harder to recognize. Nicknames are used in almost 5% of the cases. Surprisingly, the positions/roles of candidates are the least frequent mention category used in the corpus (4%).

4.3 Irony

Verbal irony is present in approximately 11% of the annotated sentences. The data shows that irony and negative polarity are proportionally distributed regarding the targets involved (Table 2). There is an almost perfect correlation between them ($r = 0.99$).

Candidate (C)	#NegCom	#IronCom
José Sócrates (C1)	766	90
M. Ferreira Leite (C2)	390	57
Paulo Portas (C3)	156	25
Francisco Louçã (C4)	171	26
Jerónimo de Sousa (C5)	109	14

Table 2. Number of negative and ironic comments

5 Main Findings and Future Directions

We showed that in our setting negative opinions tend to greatly outnumber positive opinions, leading to a very unbalanced opinion corpus (80/20 *ratio*). Different reasons may explain such imbalance. For example, in UGC, readers tend to be more reactive in case of disagreement, and tend to express their frustrations more vehemently on mat-

ters that strongly affect their lives, like politics. Anonymity might also be a big factor here.

From an opinion mining point of view, we can conjecture that the number of positive opinions is a better predictor of the sentiment about a specific target than negative opinions. We believe that the validation of this hypothesis requires a thorough study, based on a larger amount of data spanning more electoral debates.

Based on the data analyzed in this work, we estimate that 11% of the opinions expressed in comments would be incorrectly recognized as positive opinions if irony was not taken into account. Irony seems to affect essentially sentences that would otherwise be considered positive. This reinforces the idea that the real challenge in performing opinion mining in certain realistic scenarios, such as in user comments, is correctly identifying the least frequent, yet more informative, positive opinions that may exist.

Also, our study provides important clues about the mentioning of human targets in UCG. Most of the work on opinion mining has been focused on identifying explicit mentions to targets, ignoring that opinion targets are often expressed by other means, including pronouns and definite descriptions, metonymic expressions and nicknames. The correct identification of opinions about human targets is a challenging task, requiring up-to-date knowledge of the world and society, robustness to "noise" introduced by metaphorical mentions, neologisms, abbreviations and nicknames, and the capability of performing co-reference resolution.

SentiCorpus-PT will be made available on our website (http://xldb.fc.ul.pt/), and we believe that it will be an important resource for the community interested in mining opinions targeting politicians from user-generated content, to predict future election outcomes. In addition, the information provided in this resource will give new insights to the development of opinion mining techniques sensitive to the specific challenges of mining opinions on human entities in UGC.

Acknowledgments

References

Carvalho, Paula, Luís Sarmento, Mário J. Silva, and Eugénio Oliveira. 2009. "Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so easy" ;-)". In Proc. of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, Hong Kong.

Ganapathibhotla, Murthy, and Bing Liu. 2008. "Mining Opinions in Comparative Sentences". In Proc. of the 22nd International Conference on Computational Linguistics, Manchester.

Kim Soo-Min, and Eduard Hovy. 2007. "Crystal: Analyzing predictive opinions on the web". In Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague.

Krippendorff, Klaus. 2004. Content Analysis: An Introduction to Its Methodology, 2nd Edition. Sage Publications, Thousand Oaks, California.

Liu, Bing. 2010. "Sentiment Analysis: A Multifaceted Problem". Invited contribution to IEEE Intelligent Systems.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment classification using machine learning techniques". In Proc. of the Conference on Empirical Methods in Natural Language Processing, USA.

Riloff, Ellen, and Janice Wiebe. 2003. "Learning extraction patterns for subjective expressions". In Proc. of the Conference on Empirical Methods in Natural Language Processing, Sapporo.

Sarmento, Luís, Paula Carvalho, Mário J. Silva, and Eugénio Oliveira. 2009. "Automatic creation of a reference corpus for political opinion mining in user-generated content". In Proc. of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, Hong Kong.

Wiebe, Janice, Theresa Wilson, and Claire Cardie. 2005. "Annotating expressions of opinions and emotions in language". In Language Resources and Evaluation, volume 39, 2-3.

Wilson, Theresa, Janice Wiebe, and Paul Hoffmann. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis". In Proc. of the Joint Human Language Technology Conference and Empirical Methods in Natural Language Processing, Canada.