The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics

Mário J. Silva[1], Paula Carvalho[1], Luís Sarmento[2],
Pedro Magalhães[3] and Eugénio Oliveira[2]

[1] University of Lisbon, Faculty of Sciences, LASIGE
{mjs, pcc}@di.fc.ul.pt

[2] University of Porto, Faculty of Engineering, DEI - LIACC
{las, eco}@fe.up.pt

[3] University of Lisbon, Instituto de Ciências Sociais
pedro.magalhaes@ics.ul.pt

Abstract. We present the design of OPTIMISM, an opinion mining system for detection and classification of opinions about relevant political actors, regarding a particular topic of debate. The system gathers opinion-rich texts from Portuguese social media, which are then classified according to their semantic orientation and intensity. Together with the main design decisions, we present the linguistic resources for sentiment analysis in Portuguese under development and the evaluation plan. Evaluation includes comparing opinion statistics produced by the system against poll data collected purposefully for this evaluation.

Keywords: Opinion mining; Sentiment analysis; Natural Language Processing; Crowdsourcing; Social Web

1 Introduction

With the increasing availability of user-generated contents (UGC), such as blogs, Internet forums and social networks, citizens have more opportunities to express their opinions and make them available to everyone. Publicly available opinions provide valuable information for decision-making processes based on a new collective intelligence paradigm designated as *crowdsourcing* [15]. Therefore, the computational treatment of sentiment and opinions has been viewed as a challenging area of research that can serve different purposes.

Existing opinion mining systems are generally designed to deal with specific text genres and topics (for example, movie and product reviews) in English. Since specialized reviews are topic delimited, research so far has mainly focused on identifying the overall sentiment expressed in self-contained reviews (e.g. movie reviews) or the sentiment about a possible set of features that are targets of opinion (e.g. product reviews). However, more complex UGC (e.g. blog entries, opinion articles) usually express both positive and negative opinions about an open set of possible entities and topics. Hence, it is important to determine in these cases not only whether a given document, paragraph or sentence conveys a positive or negative sentiment or opinion, but also which entities are mentioned in a positive or negative way, regarding a particular issue.

This paper presents the design of OPTIMISM, an opinion mining system for detection and classification, in quasi real-time, of opinions about relevant political actors, regarding a particular topic of debate. The system gathers opinion-rich texts from Portuguese social media, which are then classified according to their semantic orientation and intensity. These include weblog posts, comments to those posts, and comments to news published in mainstream media.

The main goal of the system is to detect trends in the electoral behaviour ahead of existing polling technology by mining the social media websites, investigating how these trends correlate with those obtained using conventional polling methods [7], [18]. Recently, Google researchers have demonstrated on their *flu trends* website how they could predict influenza epidemics two weeks ahead of the existing surveillance network through the analysis of query log data [10]. We intend to attain comparable results regarding trends on political entities.

In the next section we review related work. In Section 3 we address issues related to harvesting opinionated text. Section 4 presents the set of linguistic resources being developed for sentiment analysis in Portuguese. In Section 5, we summarize our approach for learning text classifiers, which is based on a reference corpus that we are building semi-automatically. In Section 6, we present an outline of the implementation and evaluation plans for the OPTIMISM prototype. We conclude this paper by highlighting the main features of OPTIMISM.

2 Related Work

Research on opinion mining has taken three main interrelated research lines (see the comprehensive survey by Pang & Lee [25]):

(i) Development of linguistic resources for sentiment analysis, such as lexica and manually annotated corpora;

(ii) Implementation of different algorithms for text analysis and classification according to their subjectivity and semantic orientation;

(iii) Extraction of opinions from text, possibly including different types of relations with associated content.

Opinionated text has been generally classified according to sentiment polarity and intensity. Work has focused on movie and product reviews (see, among others, [23], [24] and [20]), and, more recently, on electoral behaviour [17, 21, 26]. Sentiment classifiers have been implemented using mainly machine learning based approaches [25]. Latent Semantic Analysis and Semantic Orientation – Pointwise Mutual Inclusion have been successfully applied on the construction of sentiment lexical resources [27]. However, past research experiments show that the performance of classifiers depends on multiple linguistic factors, such as text genre, topic, and, especially, the coverage and precision of linguistic information [16, 28].

Most systems for sentiment analysis make use of sentiment lexicons, containing information about the predictable polarity of words, mainly evaluative adjectives. These have been particularly studied and used as features not only to detect subjective information, but also to classify the sentiment polarity in texts [9, 13].

Simpler approaches are based on the selection of relevant lexical features, and the verification of their occurrence in a given document, which has been globally classified according to the polarity of the prevailing features. In spite of their simplicity, such approaches have shown reasonable performance in specific types of texts, such as movie reviews. However, we believe that they do not provide satisfactory results when dealing with more complex texts, involving different topics and targets of opinion. Moreover, it must be stressed that the contextual polarity of a phrase or sentence containing a sentiment word may be different from the prior polarity assigned to that word in the lexicon [29].

For example, Carvalho has shown that the meaning (and polarity) of adjectives in Portuguese depends on multiple factors, such as their position in adnominal context (e.g. *Um homem pobre / A poor man* vs. *Um pobre homem / A miserable man*), the nature of the noun to which an adjective is related (e.g. *Uma pessoa importante / An important person* vs. *Uma quantia importante / A considerable amount*) and the type of auxiliary verb with which it co-occurs (e.g. *Ele é muito inteligente / He is very intelligent* vs. *Hoje, ele está muito inteligente / Today, he is very intelligent*) [4]. She also demonstrated that some evaluative adjectives can exhibit a particular behaviour when they are included in specific constructions, such as 'cross-constructions', where the adjective fills the head of a noun phrase (e.g. *O comunista do ministro / The communist of the minister*), and exclamative constructions expressing insult (e.g. *Seu atrasado mental / You moron!*).

Regarding the selection of contents from the web for text mining, available crawlers can be classified in four major classes determined by the used harvesting strategy: (i) broad crawlers, which collect the largest amount of information possible within a limited time interval [14]; (ii) incremental crawlers, which revisit previously fetched pages, looking for changes [8]; (iii) focused crawlers, which harvest information relevant to a specific topic aided by a classification algorithm for filtering out irrelevant contents [5]; (iv) deep crawlers that also collect information relevant to a topic, but, unlike focused crawlers, have the capacity of filling forms in web pages and collect the returned pages [22].

With the advent of web 2.0 sites, such as Blogspot.com and Facebook.com, the internet domain and language-based criteria that we successfully used in the past for delimiting the Portuguese Web no longer apply (see [11], for the former, and [19], for the latter). Instead of personal web pages, political actors are now changing how they present themselves on the web, relying on hosted pages at specific applications. Many web users create web avatars that do not have a home page with true or relevant information: they simply visit popular thematic forums on the web and comment on the topic of the moment under a common nickname on a regular basis.

The opinion mining prototypes to be released by the OPTIMISM will require quasi real-time feeding with relevant Portuguese texts conveying opinions about political entities as they are published on the web. The selection of contents for mining opinions will have to follow a radically different approach.

3 Text Harvesting for Crowdsourcing

One of the main difficulties faced while crawling the first generation of the web was that contents were largely unstructured, which made their parsing hard for detecting links. On the other hand, the Portuguese web was largely organized as a collection of linked websites, each hosted under its own sub-domain of the top-level .PT Internet domain.

The selection of contents for mining opinions will have to follow a radically different approach. In the OPTIMISM crawler, we simply collect most of the data we are interested in with existing tools, by subscribing the newsfeeds associated to the relevant political actors. Their contents and meta-data are already available as data streams in the Atom or RSS2.0 syndication formats (XML). Web sites like Rssmeme.com and Pipes.yahoo.com provide support for the creation of *mashups* that can be programmed to perform some of the necessary work of aggregating newsfeed data for our demonstrator. Many news websites and blogs also enable the download of the comments by their users tied to each news or blog post.

The new challenge is finding the semantic associations among these comments, made in hundreds or thousands of independently run discussions on the same topic, having some of the users participating in several of them simultaneously. In the social web, these associations are no longer established by HTML links, but through *hashes* (twitter inherited the concept from chat forums), nicknames, and *search keywords* (where the first *hit* indirectly points to the referenced web page).

In OPTIMISM, the configuration of the automatic discovery process mainly becomes an activity of identifying the appropriate keywords for locating the relevant contents, which requires linguistic and information science skills.

4 Linguistic Resources for Sentiment Analysis

In this section, we describe the linguistic resources being developed in the scope of the OPTIMISM project, namely a sentiment lexicon and a library of syntactic-semantic patterns where polarity-bearing predicates may occur.

4.1 Sentiment Lexicon

In previous work, we explored the syntactic and semantic information described in the linguistic resources developed by Carvalho [4]. These comprise 4,250 intransitive adjectives, characterized by occurring with human subjects and having no complements. The properties considered in such resources concern, among others, the constraints imposed by adjectives respecting (i) the type of auxiliary verb with which they can co-occur, (ii) their modification by a quantifier adverb or by a morpheme of degree, (iii) their presence in specific constructions, such as *characterizing indefinite constructions* (where the adjective appears after an indefinite article, in predicative context), *cross-constructions* (where the adjective fills the head of a noun phrase), and exclamative constructions expressing insult. Furthermore, these resources describe the predicative nouns morpho-syntactically associated with each adjective (e.g. *belo/beleza; beautiful/beauty*).

We have then selected the possible polar predicates from these resources, and classified each predicate according to its predictable polarity, which may be *0*, *1*, or *-1*. These codes represent, respectively, a neutral, positive or negative semantic orientation. We are not presently assessing the potential levels of intensity exhibited by predicates from the same polarity class (e.g. *feio/horrível; ugly/horrible*), although that may be considered in the future.

The sentiment lexicon has also been enriched with new entries, collected from diverse *corpora* on the web. At the present, it is composed by a total of 6,055 intransitive predicates (more precisely, 3,533 adjectives and 2,522 names). In terms of polarity frequency, 55.5% of the predicates are classified as negatives, 21.8% as positives and the remaining 22.7% as neutral.

Future developments will also include the enlargement of the sentiment lexicon, namely in what concerns predicate verbs and multiword expressions, by exploring bootstrapping approaches [2]. We will also take into consideration particular common and proper names that may convey polarity when used metaphorically (e.g. *Ele é um verdadeiro Hitler* / *He is a real Hitler*).

4.2 Ontology of Political Entities

Entity recognition for political opinion mining in social media has several particularities, which makes it different from more generic entity recognition tasks. We are initially considering a relatively small set of pre-defined entities, including politicians and some organizations (such as political parties). In this restricted scenario, most names are unambiguous, which may suggest that simple dictionary-based recognition can lead to acceptable results.

However, a significant proportion of political mentions are indirectly made, using paraphrastic constructions, ergonyms (e.g. *o primeiro-ministro / the prime-minister*, *a líder da oposição / the leader of the opposition, o candidato do PSD à CML / the candidate from PSD for CML*). Identifying constructions of this type and associating them with their corresponding political entities is a challenging task, because they admit different lexical and syntactic variations and are prone to ambiguity. Moreover, political actors and their roles in the political scene change quickly over time (e.g. a national parliament member may become a European parliament candidate or a city mayor). Additional challenges involve dealing with absence of capitalization, acronyms, metaphoric mentions and neologisms (e.g. "Pinócrates", which results from the amalgam of the names *Pinocchio* and *Sócrates*, for referring to the current Portuguese prime-minister *José Sócrates*).

To assist the opinion extraction task, we are developing an ontology of political entities that includes the names of the political actors and corresponding variants (neologisms, acronyms, etc.), and their roles in the political scene. This ontology, which initially covers the Portuguese political environment, is being compiled semi-automatically, by mining news items using simple patterns to find possible ergonyms (usually placed in apposition to the political entity), and conflating name variations using heuristics based on lexical inclusion and edit-distance criteria. For example, we found 23 distinct mentions for the politician *Paulo Rangel* ("líder da bancada do PSD", "cabeça-de-lista social-democrata ao Parlamento Europeu", "candidato do PSD às eleições europeias", among others). We have been mining newspaper RSS feeds for about a year, to collect the most common paraphrases (presently, for more than 1300 entities). Metaphoric mentions and neologisms commonly found in UGC will be manually added to ontology, at least in a first stage. We are in the process of evaluating the ontology. We plan to use this ontological knowledge in the entity recognition module, which scans texts to identify mentions of political entities by name or paraphrase (using both exact and soft matches with the ontology contents) and performs entity resolution.

4.3 Lexico-Syntactic Rules

Opinionated messages posted in blogs, internet forums and social networks are normally short, ungrammatical and incomprehensible when taken out of context. Moreover, opinions and sentiments are mostly expressed indirectly, making use of figurative language, such as metaphors. They thus represent a hard challenge for mining approaches relying exclusively on parsers. See, for example, the fragment of a comment to a news article published in the online newspaper *Público*, which was parsed by PALAVRAS [3].

```
durão [durão] ADJ M S @SUBJ>
fez [fazer] <fmc> V PS 3S IND VFIN @FMV
muito [muito] <quant> DET M S @<ACC @>N
bem [bem] <quant> ADV @<ADVL @>A
em [em] PRP @<ADVL
fugir [fugir] V INF @IMV @#ICL-P<
de [de] PRP @<ADVS
Portugal [Portugal] PROP M S @P<
e [e] <co-fmc> <co-inf> KC @CO
ir [ir] V INF @IMV
para [para] PRP @<ADVL
um [um] <arti> DET M S @>N
super [super] <*1> <n> ADJ M S @P<
taxo [taxar] <fmc> <*2> V PR 1S IND VFIN @FMV
em [em] <sam-> PRP @<ADVL
a [o] <artd> <-sam> DET F S @>N
UE [UE] PROP F S @P<
. [.] PU <<<
```

In this case, the word *durão* was recognized as an adjective, instead of a proper noun (*Durão Barroso*), because it is written in lowercase. On the other hand, the prior negative noun "tacho" (*approx.* "job for the boys") was analyzed as a verb (*taxar / to tax*) due to a common "Internet-accepted variation" ("*ch*" → "*x*").

We believe that robust opinion detection/classification in UGC can only be achieved by using text classification techniques exploring the rich set of features that may be extracted from text and other fine-grained linguistic resources. Some of these features may consist of lexically-derived information, such as n-grams or information about the prior-polarity of words, while others may be identified using lexico-syntactic patterns to detect typical opinion-bearing structures. Hence, in parallel to lexicon development, we are creating high precision lexico-syntactic rules. We started by confining the analysis to an ensemble of typical syntactic constructions where predicative adjectives and nouns, such as the ones described in the sentiment lexicon, may occur. Below, we illustrate some elementary adjectival predicative constructions including a political entity (PE):

(1) PE Vasp* Prep$^{?}$ Vcop^{1} Adv* Adj
(e.g. *Sócrates continua a ser muito convencido*$_{neg}$→S_{neg})
(2) PE Vasp* Prep$^{?}$ Vaux^{1} Artind Adj$_{pos|neut}$ Adj$_{neg}$
(e.g. *Sócrates é um bom*$_{pos}$ *mentiroso*$_{neg}$→S_{neg})
(3) PE Advneg Vasp* Prep$^{?}$ Vaux^{1} Adv* Adj
(e.g. *Sócrates nunca foi desonesto*$_{neg}$→S_{pos})
(4) PE deixar de Vaux^{1} Adv* Adj
(e.g. *Sócrates deixou de ser autoritário*$_{neg}$→S_{pos})
(5) PE Vasp* Prep$^{?}$ Vaux^{1} Adv* Adj, mas Adv* Adj
(e.g. *Sócrates é arrogante*$_{neg}$, mas *competente*$_{pos}$→S_{pos})
(6) Artdef Adj$_{pos|neut}$ do|da|dos|das PE...
(e.g. *O inteligente*$_{pos}$ *do Sócrates...* →S_{neg})

In construction (1), the adjective occurs in a predicative context, i.e., it relates a PE with their subject through a copulative verb (*Vcop*), which may be preceded by an aspectual verb (*Vasp*). Moreover, the adjective may be modified by an adverb (*Adv*). The application of this regular expression makes it possible to extract the sentences that match this pattern, and classify them according to the polarity assigned to adjectives in the lexicon.

However, sentences often include more than one lexical unit assigned to opposite polarities in the lexicon, as illustrated in construction (2). As a result, it is crucial to identify which is the predicate of the sentence, in order to classify it. Different types of negation are also considered in syntactic-semantic rules, as illustrated in constructions (3) and (4). The sentences matching those patterns must be classified as positive or negative, by reversing the polarity value assigned to the predicate they refer to. This implies that they must be classified as negative if containing a positive predicate, and vice-versa.

Lexico-syntactic rules also account for the possibility of co-occurrence in the same sentence of two or more predicates presenting similar or opposite polarities. For instance, the subject of construction (5) is modified by two predicates, which are combined in a coordinate structure, by means of the *mas* (*but*) adversative conjunction. In this case, we classify the sentence according to the polarity code exhibited by the last element of the coordination.

In some cases, the prior polarity assigned to a predicate at lexical level can be changed in particular contexts, such as the case of construction (6).

5 Learning Opinion Classifiers

Practical opinion classification cannot be performed using only a set of predefined lexico-syntactic rules [6]. To allow generalization, and thus increase recall in opinion classification, it is crucial to train an automatic text classifier. For this purpose, we are currently developing a reference corpus for the training and evaluation of opinion classification procedures.

5.1 Reference Corpus

We started by collecting opinion-rich data from the web site of one of the most read Portuguese newspapers. We obtained a collection of 8,211 news and linked comments to that news posted by on-line readers. This collection covers a period of 5 months (November 2008 to March 2009). It includes about 250,000 user posts, containing approximately 1 million sentences. We are now proceeding with the annotation of this collection in order to develop the reference corpus. We will identify mentions to political entities in the comments (both by name and by paraphrastic or anaphoric mention) and detect whether there is an expressed positive or negative opinion about those entities at the sentence level.

At present, we are focusing on a small set of popular political entities occurring in specific constructions, including those described in Section 4.3, in order to find high precision candidate sentences expressing positive and negative opinions. Such candidate sentences are being manually validated, and the remaining sentences of the comments mentioning the same entity are being automatically classified. This procedure is based on the assumption that non-recognized opinionated constructions expressed in a particular comment about a specific entity are consistent with the opinions previously found by rules in the rest of the comment. This enables quick annotation of the collection about certain political entities, which includes opinionated sentences with a large diversity of structures. Some of these would not be easily detected by linguistic rules.

Results obtained so far show that the performance of lexico-syntactic rules depends on the polarity conveyed in recognized text. In the case of expressed negative opinions, the precision rate is approximately 90%. For positive opinions expressed in text, the precision decreases to 60%. In both cases, we achieved a very low recall, which is not surprising due to the small number and strict type of syntactic patterns explored in this work.

Our experiments also demonstrate that it is possible to propagate the opinions found by the lexico-syntactic patterns to other sentences of the user comments mentioning the same entity, increasing the number and diversity of annotated sentences. Again, propagation success is higher for negative opinions (almost 100%) than for positive opinions (around 75%).

5.2 Feature Selection and Classification Algorithms

The challenge for correct classification lies in selecting the correct set of features for describing opinions in text. Some of the relevant features for opinion classification are associated with the presence in text of elements listed in the sentiment lexicon. However, there are relevant clues about opinion in features related to (i) punctuation usage (e.g. heavy punctuation); (ii) graphic marks (e.g. full capitalization); (iii) distance of elements to the entity at stake (e.g. immediately before/after entity); (iv) POS information, and (v) more refined syntactic and semantic information. The best feature selection of features will be investigated by evaluating the performance of classification over the reference corpus with different machine learning algorithms [12].

So far we have made initial experiments in a more constrained setting. We manually annotated a selected set of about 2,000 short news (title plus 1-3 sentences) gathered from newspaper RSS feeds and covering either a negative or positive event associated to two political entities: the prime-minister and the leader of the opposition of Portugal. We trained an SVM classifier for predicting if unseen news are about either a positive or negative event associated to one of the two. We experimented several feature sets, including bag-of-words, n-grams plus relative positions, prior polarity of words and POS. Results show that we are able to achieve a precision rate over 90% and an F-measure over 80% in predicting whether news are positive or negative for several of the possible feature combinations. Although UGC represents a much greater challenge, and despite the fact that we are considering a scenario with fewer restrictions in OPTIMISM, we believe that these results support the feature exploration strategy we are following.

6 Implementation and Evaluation Plans

We plan to release software prototypes capable of processing Portuguese texts conveying opinions about relevant political entities as they are published on the web for user evaluation. We have adopted an agile methodology organized in short sprints that add/modify functionalities as a result of observing new potential uses, alternative user interfaces or new characteristics of the opinion mining and web crawling modules of the system. The software components will be continuously optimized, as the training corpora improve and we obtain user feedback.

We started by defining a simple database model for storing the collected news items and comments (and metadata). A focused web crawler populates the database and the opinion mining software then fetches sets of news items for annotation from the database. The OPTIMISM software will also include a library of components for computing statistics on opinion data and plotting these statistics as charts and trends. The actual user interface design on both prototypes will depend on the characteristics of the obtained data, the methods used to process such data, and observation of user access patterns.

The evaluation of the OPTIMISM crawler will focus on measuring its capability to re-discover, from a set of initial seeds, known relevant blogs and Web2.0 from a known list to be compiled by the political scientists in the team, while minimizing the number of irrelevant blogs and comments to the task at hand.

Evaluation of the system will be performed through periodic user satisfaction surveys. We will also need to ensure that the perception on the quality and usefulness of the mining software is not hampered by a poorly designed user interface.

Planned experiments also include exploring the relationship between sentiments towards political entities as measured by the prototype and those obtained in political opinion polls. On the one hand, polls to be conducted by the Catholic University's Poll Centre will include questions designed to detect sub-samples of social web users, enabling the matching of data generated by the prototype with that obtained among representative samples at the same time periods. On the other hand, trends in all published public opinion polls and those obtained through the prototype with be analyzed both through conventional time series techniques and wavelet decomposition techniques appropriate for analyzing the relationship between two signals and the role of critical events [1].

7 Conclusion

OPTIMISM is a system for performing real-time opinion mining about political entities over the enormous wealth of user-generated content being currently produced. This system differs from most previous opinion mining projects in multiple aspects. It aims at detecting opinions about specific political entities regarding an open set of topics or issues. It is intended to explore highly reactive and short-lived user-generated content such as social web sites' status messages and comments posted on on-line newspapers, which represent a great challenge due to their short, and often non-grammatical, nature. For content selection, we are planning to explore deep crawling through interaction with global search engines (that perform broad crawls) for locating new potentially relevant content, but we incorporate aspects of incremental and deep crawling for obtaining the data.

We are currently using a novel semi-automatic strategy for building a reference corpus for training text classifiers, which mines newspaper comments with high-precision lexical-syntactic patterns to gather a diverse set of examples of opinionated text. Such corpus will be used to explore feature selection policies to be used by text-classification algorithms.

Finally, the evaluation plan includes contrastive analysis of observation data derived from the prototype's use against true political opinion poll data.

Acknowledgments:

This work was partially supported by grants SFRH/BPD/45416/2008 and SFRH/BD/23590/2005 from FCT-Portugal, co-financed by POSI.

References

1. Aguiar-Conraria, L., Magalhães, P.C.: Dynamical decomposition of political time-series, paper prepared for the American Political Science Association 2009 meeting.
2. Banea, C., Mihalcea, R., Wiebe, J.: A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco (2008)
3. Bick, E.: The Parsing System "Palavras": *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Dr.phil. thesis. Aarhus University. Aarhus University Press. Denmark (2000)
4. Carvalho, P.: *Análise e Representação de Construções Adjectivais para Processamento Automático de Texto. Adjectivos Intransitivos Humanos*. PhD Thesis, University of Lisbon (2007)
5. Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., Kleinberg, J.: Mining the Web´s link structure. In *Computer*, 32(8):60.67 (1999)
6. Dasgupta, S., Ng, V.: Mine the Easy, Classify the Hard: Experiments with Automatic Sentiment Classification. To appear in *Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, 2009.
7. Drezner, D, Farrell, H.: Introduction: Blogs, Politics and Power, In *Public Choice*, 134, 1-2, pp. 1--13 (2008)
8. Edwards, J., McCurley, K.S., Tomlin, J.A.: An adaptive model for optimizing performance of an incremental web crawler. In *International World Wide Web Conference*, pp. 106–113 (2001)
9. Esuli, A., Sebastiani, F.: Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, EACL'06*, Trento, Italy (2006)
10. Ginsberg J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data, *Nature* (19 Nov 2008), doi: 10.1038/nature07634, Letters to Editor (2008)
11. Gomes D., Silva M.J.: The Viúva Negra crawler: an experience report. In *Software: Practice and Experience (SPE)* 2(38), pp. 161–168 (2008)
12. Guyon, I., Elisseff, A.: An Introduction to Variable and Feature Selection**.** In *Journal of Machine Learning Research* 3, pp. 1157--1182. (2003)
13. Hatzivassiloglou, V., McKeown, K.: Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, Madrid, Spain (1997)
14. Heydon, A., Najork, M.: Mercator: A scalable, extensible web crawler. In *World Wide Web*, 2(4), pp. 219–229 (1999)
15. Howe, J.: The Rise of Crowdsourcing, *Wired*, Vol. 14, No. 6 (2006)
16. Kim, S., Hovy, E.: Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia (2006)
17. Kim, S, Hovy, E.: Crystal: Analyzing predictive opinions on the web. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2007)
18. Magalhães, P.: Pre-Election Polls in Portugal: Accuracy, Bias and Sources of Error, 1991-2004, *International Journal of Public Opinion Research* 17 (4), pp. 399--421 (2005)

19. Martins, B, Silva, M.J.: Language Identification in Web Pages. In *Proceedings of ACM-SAC-DE, the Document Engineering Track of the 20th ACM Symposium on Applied Computing*, pp. 764--768, Santa Fe, Mexico (2005)
20. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputations on the Web. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 341--349 (2002)
21. Mullen, T., Malouf, R.: A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 159--162 (2006)
22. Ntoulas, A., Zerfos, P., Cho, J.: Downloading textual hidden web content through keyword queries. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pp. 100–109, ACM Press, NY, USA (2005)
23. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd ACL*, Barcelona, Spain (2004)
24. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, Philadelphia, USA (2002)
25. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis, In *Foundations and Trends in Information Retrieval* 2(1-2), pp. 1–135 (2008)
26. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 327–335 (2006)
27. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. In *ACM Transactions on Information Systems*, 21(4), pp. 315–346 (2003)
28. Whitelaw, C., Garg, N., Argamon, S.: Using Appraisal groups for sentiment analysis. In *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM 2005)*, Bremen, Germany (2005)
29. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. In *Computational Linguistics* 35(3), pp. 1-34 (2009)