LEGi: Context-aware lexicon consolidation by graph inspection

Article · March 2014

DOI: 10.1145/2554850.2554916

6 authors, including:

Giovanni Sá
Federal University of São João del-Rei
4 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Thiago Silveira
Federal University of São João del-Rei
9 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Rodrigo Chaves
Federal University of São João del-Rei
3 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Fernando Mourão
University of Minnesota Twin Cities
52 PUBLICATIONS **266** CITATIONS
SEE PROFILE

LEGi: Context-Aware Lexicon Consolidation by Graph Inspection*

Giovanni Sá[1], Thiago Silveira[1], Rodrigo Chaves[1],
Felipe Teixeira[1], Fernando Mourão[2], Leonardo Rocha[1]

[1]Universidade Federal de São João Del Rei
Computer Science
São João Del Rei, Brazil
{giovannisa, tssilveira, rachaves, fcteixeira, lcrocha}@ufsj.edu.br

[2]Universidade Federal de Minas Gerais
Computer Science
Belo Horizonte, Brazil
fhmourao@dcc.ufmg.br

ABSTRACT

The value of subjective content available in Social Media has boosted the importance of Sentiment Analysis on this kind of scenario. However, performing Sentiment Analysis on Social Media is a challenging task, since the huge volume of short textual posts and high dynamicity inherent to it pose strict requirements of efficiency and scalability. Despite all efforts, the literature still lacks proposals that address both requirements. In this sense, we propose LEGi, a corpus-based method for consolidating context-aware sentiment lexicons. It is based on a semi-supervised strategy for propagation of lexicon-semantic classes on a transition graph of terms. Empirical analyses on two distinct domains, derived from Twitter, demonstrate that LEGi outperformed four well-established methods for lexicon consolidation. Further, we found that LEGi's lexicons may improve the quality of the sentiment analysis performed by a traditional method in the literature. Thus, our results point out LEGi as a promising method for consolidating lexicons in high demanding scenarios, such as Social Media.

Categories and Subject Descriptors

H.3.m [**Information Storage and Retrieval**]: Miscellaneous; I.2.7 [**Natural Language Processing**]: Text Analysis

General Terms

Algorithms, Evaluation

Keywords

sentiment analysis, sentiment lexicon, context-aware

*This work was partially supported by CNPq, CAPES, Fapemig, and INWEB.

SAC'14 March 24-28, 2014, Gyeongju, Korea.

1. INTRODUCTION

Social media has emerged as an important environment in which people publish their opinions about distinct topics on the Web, providing an unprecedented source of information for modeling and understanding user behaviors. Not surprisingly, a growing number of efforts aim to adapt traditional computational analysis to this new scenario. Given the practical value of such subjective content, particular attention has been given to **Sentiment Analysis** on Social Media, that is, the automatic extraction and identification of subjective information from textual data [10, 11, 17].However, performing Sentiment Analysis on Social Media is even more challenging than on other scenarios, such as product reviews, due to the huge volume of short textual posts and high dynamicity of data generation common to stream scenarios [2].

In this work, we address the challenge of performing effective and efficient Sentiment Analysis on Social Media streams. Indeed, this goal has been previously pursued in the literature, mainly, through unsupervised lexicon-based Sentiment Analysis methods [9, 14]. Basically, these methods build lexicons, structures that assign a polar class (i.e., positive, negative or neutral) to each term, and use this information to classify each document w.r.t. its polar class. Hence, the quality of sentiment analysis becomes directly related to the classification's precision of these lexicons. Although manually built lexicons exhibit good results, they are labor intensive and, consequently, not appropriate for high demanding scenarios. Thus, recent studies build automatically lexicons based on document corpus, analyzing more efficiently different domains [6, 10, 14]. Further, these corpus-based approaches classify the sentiment of each term according to each domain, since such sentiment might differ among domains. For instance, the word 'fun' may have a positive connotation in the context of TV series and a negative one in political contexts. However, these methods still are computationally prohibitive for handling social media content.

Therefore, we propose LEGi, an efficient corpus-based method for consolidating context-aware sentiment lexicons. LEGi is based on a semi-supervised strategy for propagation of lexicon-semantic classes on a transition graph G of terms. The set of distinct terms observed in the analyzed document corpus D and their synonyms and antonyms, present in a thesaurus, define the node set of G. In turn, there are two different types of edges in G. The first type determines that two terms are synonyms, whereas the other determines they are antonyms. Starting from few seed terms, previously classified w.r.t. their lexicon-semantic class by specialists of

the evaluated domain, at each iteration, we propagate the sentiments of all classified terms to their adjacent nodes in G. If we reach a node by a synonym's edge, we propagate the original class to this node, otherwise we propagate the reverse class. We change the class of a node whenever most of the propagation that reach it define a different class. This process continues until no node of G changes its class or a maximum number of iterations has been reached.

Empirical assessments on LEGI demonstrate its applicability in two real domains composed by Twitter posts. The first domain comprises 20 USA TV series (entertainment) and the second one refers to the 2012 USA presidential election (political). In each domain, we contrast the lexicon generated by LEGI against four other lexicons provided by well-established lexicon consolidation methods in the literature [16, 12, 6, 1]. These analyses show that, besides achieving the best approximation to the gold standard lexicon of each domain, in which all terms were manually classified accordingly to the domain of analysis, LEGI presented a small execution time. Further, we found that LEGI's lexicons may improve the quality of the sentiment analysis performed by a traditional method in the literature [17].

2. RELATED WORK

In the past few years, we have observed a growing interest on the field of Sentiment Analysis, whose goal is to extract subjective information from textual data [10, 17]. Such interest stems from the consolidation of social media as a new, rich and huge source of subjective information about users. However, several studies have found that traditional methods of sentiment analysis, employed in scenarios such as product reviews, are not suitable for this new scenario. Most of these methods cannot handle a huge volume of data, since they are costly supervised classification algorithms based on Machine Learning. Besides traditional features used in Automatic Document Classification (e.g., word occurrence), they usually consider features extracted by Natural Language Processing methods, such as tuples of terms or syntactic classes of terms. Further, these classification models should be continuously updated due to the high dynamicity of social media scenarios. Thus, unsupervised methods are assuming an important role on the pursuit of effective and efficient approaches for sentiment analysis in social media content [9, 14].

Most of the unsupervised sentiment analysis methods are based on sentiment lexicons and have usually two distinct steps. In the first one, the consolidation of a lexicon is performed. Based on such lexicon, the second step focuses on identifying the sentiment of each distinct document. Hence, the consolidation of high quality lexicons is crucial for this kind of method. In turn, as presented in [4], studies on the consolidation of lexicons can be divided into three main categories. The first one refers to methods based on human annotations [16, 12] on which terms are classified manually. Despite providing good lexicons, these methods are labor intensive and, therefore, unfeasible for social media applications. The second category comprises dictionary-based methods [1, 7], that identify the sentiment of a term from other semantically related terms in a specific dictionary, e.g., WordNet. These methods are usually computationally efficient, however do not take into account the application context. Finally, the third category is known as corpus-based methods [6, 14, 10] on which a context-dependent sentiment is constructed, defining the sentiment of terms according to relations between terms observed in a corpus. The lexicons built by these methods might be as good as the manual construction with a reduced computational cost. However, the time required for this sort of analysis is still prohibitive for scenarios like social media applications.

In this paper, we present LEGI, a new method that belong to the third category, since it takes into account the application domain in the lexicon consolidation process. Adopting a semi-supervised strategy, LEGI becomes computationally suitable to high demand scenarios, such as social media.

3. SEMI-SUPERVISED SENTIMENT LEXICON CONSOLIDATION

In this section, we describe LEGI, our method to consolidate a context-aware semantic lexicon. Such as the previous lexicon-based proposals [9, 17, 15], LEGI assumes that there are distinct classes of words that play specific roles on the consolidation of subjective sentences [8].

This lexicon-semantic classification is usually performed regardless of the analyzed domain, assigning a single *prior* class to each term [3]. However, recent works have pointed out that some terms may take different sentiments on distinct domains [10]. Thus, LEGI classifies each term according to each domain of analysis, described through huge corpus of textual data.

We designed LEGI taking into account four main requirements of analysis. The first one is *instantiability*, which we define as the capability of a method be easily instantiated to distinct domains. The second requirement refers to *coverage*, which is the percentage of distinct terms of a domain classified w.r.t. lexicon-semantic classes. Our third requirement comprises *computational efficiency*, since our goal is to analyze large volumes of data in short periods of time. Finally, the fourth requirement concerns *classification efficacy*, measured through the percentage of classified terms assigned to the correct lexicon-semantic class. LEGI aims to achieve all these requirements simultaneously, which comprises a challenging task given the not aligned goals usually posed by them.

Table 1: Subset of lexicon-semantic classes observed in the English vocabulary.

Class	Transformation	Examples
Positive	Makes positive the sentiment of a sentence.	good, nice, amazing
Negative	Makes negative the sentiment of a sentence.	terrible, ugly, bad
N**E**utral	Does not change the sentiment of a sentence.	most of substantives

In this sense, we propose LEGI as a semi-supervised algorithm of lexicon-semantic classes propagation on a transition graph of terms. The propagation starts with a few distinct seed terms, which have been manually classified w.r.t. pre-defined lexicon-semantic classes by, for instance, specialists of each evaluated domain. Specifically, we adopted the pre-defined set of lexicon-semantic classes presented in Table 1, which constitutes the subset of three polar morphological classes, proposed in [16], most frequently adopted by semantic-lexicon studies in the literature [6, 14, 10, 13].

LEGI starts by building a transition graph G among terms. This graph is defined based on the set T of distinct terms observed in a finite set of opinative documents D and on

external information of synonyms and antonyms (i.e., a thesaurus) for the language used in D. Thus, we define as nodes of G all terms of T as well as their synonyms and antonyms. We also define two types of edges in G. There is a *s-edge* between two terms, if they are synonyms in the thesaurus and an *a-edge* if they are antonyms. Once the graph has been constructed, the next step is to classify manually a subset $S \subset T$ of terms (i.e., the seed set) regarding the adopted lexicon-semantic classes. We assume that S is much smaller than T (i.e., $S \ll T$) and its terms are those with highest frequency of occurrence in D.

Algorithm 3.1 Propagation of lexicon-semantic classes

```
1: function DEFINESENTIMENTLEXICON(G, V, U, Φ, γ, δ)
2:     classifiedTerms ← V
3:     unknownClasses ← U
4:     auxilarySet ← {}
5:     propag ← 0
6:     while (classifiedTerms ≠ auxilarySet)&(propag < δ) do
7:         auxilarySet ← classifiedTerms
8:         for t_i ∈ classifiedTerms do
9:             SentimentClass ← retrieveSentimentClass(t_i)
10:            neighbors ← getNeighbors(t_i, G)
11:            for all t_j ∈ neighbors do
12:                edgeType ← retrieveEdgeType(t_i, t_j, G)
13:                if edgeType ≡ a-edge then
14:                    SentimentClass ← invertSentiment(SentimentClass)
15:                propagations.push(t_j, SentimentClass)
16:        for all t_i ∈ propagations do
17:            numPropagations ← getNumTuples(propagations, t_i)
18:            majorClass ← getMajorClass(propagations, t_i)
19:            majorClassFreq ← getClassFreq(propagations, t_i, majorClass)
20:            confidence ← majorClassFreq/numPropagations
21:            if (numPropagations > Φ)&(confidence > γ) then
22:                SentimentLexicon[t_i] ← majorClass
23:                classifiedTerms ← classifiedTerms ∪ t_i
24:                unknownClasses ← unknownClasses − t_i
25:        propag++
       return SentimentLexicon
```

Basically, the consolidation of the sentiment lexicon consists in propagating classes from the seed nodes in G, as described by Algorithm 3.1. First, we define two distinct sets of terms. One of them represents the set V of all terms with known class (i.e., initially $V = S$) and the second set U comprises the remaining terms, which are assigned to the neutral class (i.e., $U = T - V$). The propagation is an iterative procedure (lines 6 to 25). At each iteration, all nodes t_j adjacent to each node $t_i \in V$ are checked. If the edge that links t_i to t_j is a *s-edge*, then the class of t_i is propagated to t_j. If the edge is an *a-edge*, then we propagate the reverse sentiment of t_i to t_j, wherever applicable (positive-negative or negative-positive). At the end of each iteration (lines 16 to 24), there is a decision step that determines whether we should change the class of a node t_j. This decision is based on all propagation paths that reach t_j and two input parameters. The first one is the support Φ that determines the minimum number of neighbors, who have propagated their classes to t_j, necessary to reclassify t_j. The second parameter is the minimum confidence γ, which is the relative frequency of the most frequent lexicon-semantic class among those propagated to t_j (i.e., the major class). If these two criteria are satisfied, the class of t_j is updated to the major class. Then, t_j is removed from the set U and inserted into the set V. In the next iteration, this process is repeated with the new set V and the process continues until V does not change anymore or a maximum propagation distance δ, given as parameter, is reached. Finally, all terms not reached by this propagation process are assigned to the neutral class.

Through Algorithm 3.1, we observe that LEGI covers the four aforementioned requirements. We address *instantiability* by adopting a restricted set of seed terms, manually classified according to each domain, and a semi-supervised strategy, reducing the amount of manual effort required. Further, the propagation of classes allows us to address *coverage* of terms. In turn, *efficiency* is ensured by adopting a simple majority voting process. Finally, we tune *efficacy* through three input parameters: maximum distance of propagations (δ), support (Φ) and confidence (γ). Such parameters should be carefully selected according to each domain, since they may vary with the analyzed set of terms.

4. EXPERIMENTAL EVALUATION

4.1 Datasets

We performed all empirical analyses using two quite distinct data samples collected from *Twitter*, a worldwide popular microblogging service, corresponding to two different domains (entertainment and politics). The first dataset (TV Series) is composed of *tweets* related to 20 American TV Series with global audience, such as *The Big Bang Theory* and *Two and a Half Men*, among others. The second dataset (USA Elections) comprises *tweets* related to the 2012 USA presidential elections. We gathered both collections through the Twitter search API by using hashtags and common words related to names or acronyms of representative entities existing in each domain. More specifically, we used as representative entities all the 20 distinct series observed in TV series, while only the candidate Barack Obama was considered in USA Elections. Further, we removed from the textual data all punctuations, special characters, repetitive patterns (e.g., LOOOOVE becomes love), as well as we converted all letters to lowercase. Also, words with less than three or more than 12 letters, URLs, emotional signals and quotes to users were removed from both datasets. Table 2 presents detailed information about each resulting dataset.

Table 2: Dataset descriptions.

Datasets	# Tweets	Positive Tweets	Negative Tweets	Neutral Tweets	Distinct Terms
TV Series	3,117	27.94%	4.52%	67.53%	4,503
USA Elections	1,353	20.33%	30.88%	48.79%	3,635

Aiming to provide inputs for the evaluation metrics, each *tweet* from both of the datasets was manually classified, as well as each distinct term observed in each collection w.r.t. its lexicon-semantic class, defining a gold standard tweet set and a gold standard lexicon respectively. In both cases, each tweet or term was classified into one of three distinct classes (positive, negative and neutral) by five different people and, for each tweet or term, the class with majority votes was taken as the gold standard. In the few cases that there were indecision, the tweets and terms were classified as neutral. Table 2 presents the class distribution for tweets, while Table 3 shows the class distribution in the gold standard lexicons. We observe that neutral terms represent approximately 90% of all terms, in both datasets. This behavior

is not surprising. Besides the fact that human language is composed mostly by neutral terms, *tweets* typically comprise short sentences whose sentiment tends to be defined by few distinct terms.

Table 3: Distributions of lexicon-semantic classes on gold standard lexicons.

Dataset	Positive Terms	Negative Terms	Neutral Terms
TV Series	4.3%	2.0%	93.7%
USA Elections	4.9%	8.4%	86.7%

4.2 Analysis of Parameters

In this section, we evaluate the impact of each parameter required by LEGI on the four requirements discussed in Section 3: **instantiability**, **coverage**, **efficiency** and **efficacy**. As previously discussed, LEGI requires four parameters: the initial seed set (S), the maximum propagation distance (δ), the confidence (γ) and the minimum support (Φ). In order to better understand the relevance of each parameter, as well as the correlation among distinct ones, we defined fine-grain discrete intervals of analysis and performed a factorial experiment. The resulting lexicons of several parameter combinations were contrasted to the gold standard lexicon. First, we discretize confidence into intervals of size 0.1, from 0 to 1. Also, we vary the size of the initial seed set from 1 to 100 by using distinct subsets of the most frequent terms found in each gold standard lexicon. Values greater than 100 are not evaluated, since it is crucial to construct semantic lexicons minimizing the manual effort (i.e., manual classification of terms). For minimum support, we vary the values from 2 to 40. However, it was observed that values greater than 10 do not bring significant differences in the quality of sentiment lexicons. Finally, we vary the maximum propagation distance values from 1 to 10, but there was no difference on the quality of semantic lexicons for values greater than 6. Thus, we present a pairwise analysis of parameters, considering just the intervals in which the differences were significant and fixing the two remaining parameters to the following values: $\delta = 3$, $\Phi = 3$, $\gamma = 0.8$ and $|S| = 100$, which correspond to values that exhibited the best results for the lexicon construction.

The analysis starts by the coverage requirement, as shown by Figure 1 (a). Coverage was measured through the *Global Recall* of the gold standard lexicon, which accounts for the percentage of the gold standard lexicon properly classified by LEGI. Figure 1 (a) shows that, for both collections, more than 70% of each gold standard lexicon can be classified starting from small sets of seeds (i.e., less than 100 terms). The best combinations of parameters presented coverages of 75% and 83% for TV Series and USA Elections, respectively. Moreover, using high values of minimum support is not recommended because they degrade coverage. Furthermore, we observe that high confidence values degrade coverage. Finally, we found that the greater the distance of propagation, the higher the resulting coverage, since by increasing this distance, the number of reached terms tends to grow exponentially in this type of graphs.

Efficacy was evaluated through the well-known $Macro\text{-}F1$ metric that corresponds to the average of per class $F1$, which is calculated as the harmonic mean between precision and recall of each class [5]. $Macro\text{-}F1$ exhibits a greater capability of handling skewed class distributions than other

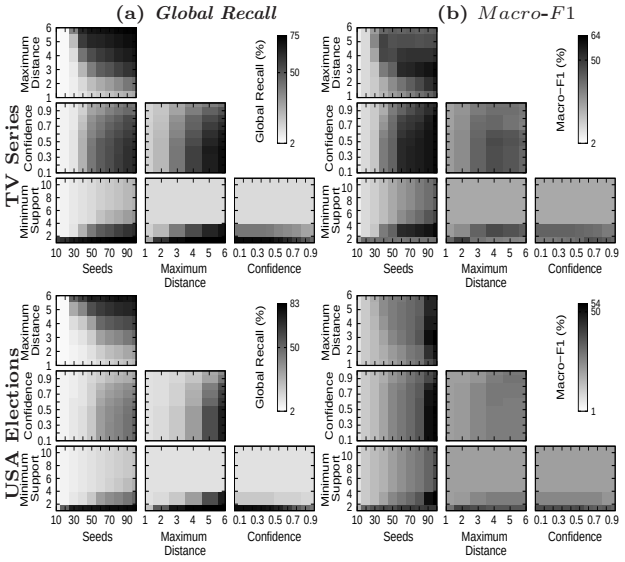

Figure 1: Qualitative analysis of parameters in the LEGI's sentiment lexicon construction.

traditional metrics, such as *Accuracy*. By analyzing Figure 1 (b), it can be seen that, for sake of efficacy, neither higher values of minimum support nor lower ones should be used. High support values filter out almost all terms and, consequently, most of them remain as neutral. On the other hand, low support values do not filter out noises and unreliable propagation. In general, this noise stems from originally neutral terms that have some synonyms or antonyms different from neutral and, consequently, are assigned to another class. Analogously to the coverage evaluation, confidence has little impact on efficacy. Finally, it is possible to note that, despite high maximum propagation distances provide high coverages, these distances are not suitable regarding efficacy, because they increase the number of misclassifications. This problem stems from the absence of the transitivity property on the meaning, given that the synonym of a synonym may not have a meaning close to the original one.

Taking into account instantiability, we analyze the seed set behavior on Figure 1. We note that, as expected, the larger the set of initial seeds, the higher the coverage and effectiveness of the generated lexicon. However, for the TV Series dataset, an initial set of only 50 seeds was enough for achieving high coverage and efficacy levels. On the other hand, the USA Elections dataset required a larger initial set with at least 90 seeds. This behavior is related to levels of subjectivity and specificity inherent to the vocabulary of each domain. For instance, in TV Series positive terms such as *good* and *funny* are quite common, sharing the same sentiment and positive synonymous or negative antonyms with many domains. In USA Elections instead, the most frequent positive terms are very specific to this domain, such as *vote* or *reelect*. Hence, the number of seeds required for properly identifying the lexicon-semantic class of terms in the political domain tends to be high, since several of its terms are usually neutral in other domains. Nevertheless, the number of seed terms required for achieving good coverage and efficacy levels represents less than 3% of the vocabulary size in both collections. This fact evinces that LEGI can be easily instantiated on distinct domains.

Table 4: Comparison of effectiveness among distinct lexicon consolidation methods into two distinct domains. The best results for each metric are shaded. We found that LEGI presented the best results.

Method	TV Series				USA Elections			
	F1-P (%)	**F1–N** (%)	**F1–E** (%)	**Mac-F1** (%)	**F1-P** (%)	**F1–N** (%)	**F1–E** (%)	**Mac-F1** (%)
MPQA	45.61	73.49	15.83	44.98	35.20	52.24	39.77	42.40
INQ	48.51	79.25	16.99	48.25	35.27	57.03	29.42	40.57
SentiWordNet	24.06	80.92	23.69	42.89	16.75	**62.40**	30.76	36.64
BL	49.49	78.31	21.21	49.67	37.46	50.96	**40.30**	42.90
LEGI	**58.17**	**84.36**	**25.64**	**56.06**	**38.69**	52.10	**40.30**	**43.79**

Finally, we evaluate LEGI's efficiency, since our goal is to handle huge volumes of data in short time periods, providing a method suitable to highly dynamic scenarios, such as streams of data. We assess this efficiency requirement by conducting a worst-case asymptotic analysis of complexity on LEGI, allowing a broader and more robust evaluation. The worst case happens when the number of initial seeds is equal to one and each term has only a unique and distinct neighbor, defining a linear list of size equal to the vocabulary size. Hence, at each iteration, just one term is classified and the time complexity required for the sentiment lexicon convergence is $O(k^2)$, where k refers to the number of distinct terms being classified. Nevertheless, this worst-case scenario is quite unlikely in practice and the average number of neighbors per term tends to be significantly higher than one. Further, empirical evaluations corroborated this last observation. For any set of seeds used in both datasets, the experiments showed that only 10 iterations were required for the lexicon convergence. Moreover, by measuring the execution time of LEGI for generating a specific lexicon for each dataset, considering a thesaurus with 71,000 words, we have found times lower than 9 seconds. The execution time was calculated as the average of 10 executions, varying the seed set, on an Intel Core i5 1.8 GHz machine with 4GB of RAM and running Ubuntu 12.04.

An immediate question raised by the foregoing discussion is how to obtain, in a practical manner, a *'proper parameter calibration'* for LEGI? As the previous analyses use both a gold standard lexicon, whose consolidation is expensive, and a costly exhaustive search for the best combination of parameters, we should point out feasible strategies to set these parameters, enhancing the applicability of our method. A simple strategy would be to extract a small subset of the most frequent terms present in the analyzed dataset and build the gold standard lexicon for this subset. Then, an almost exhaustive experiment could be performed to find the parameter values that produce a resulting lexicon closest to the gold one. Next, these parameters could be used to determine the sentiment lexicon on the original dataset. Thus, we highlight, as immediate future work, the proposal and evaluation of distinct feasible parameter calibration strategies, as well as further changes in LEGI in order to operationalize it for data streams scenarios.

4.3 Comparison of Lexicons

The actual effectiveness of LEGI in real domains comprises the main dimension of evaluation. In this section, we perform such evaluation by contrasting the lexicons defined by LEGI in our collections against four baseline lexicons commonly used in the literature [16, 12, 6, 1]. These baselines belong to the three categories mentioned in Section 2. The first baseline, Multi-Perspective Question Answering (MPQA), was originally proposed in [16] and belongs to the first category. MPQA has 8.221 words manually classified, with 2.718 positive, 4.912 negative and 571 neutral words. This lexicon also has the grammatical class of each word, as well as its level of importance in a sentence. The second baseline, Harvard General Inquirer (INQ) [12], also belongs to the first category. It is a lexicon that has words manually classified syntactically, semantically and grammatically. INQ has 11.788 words, 1.915 positive, 2.291 negative and 7.582 neutral words. As our third baseline, we evaluate the SentiWordNet [1] lexicon (version 3.0) that belongs to the second category and assigns three numerical values for each word, which corresponds to its levels of positivity, negativity and neutrality. SentiWordNet has 148.610 classified words on which, considering that the highest value assigned to a word corresponds to its class, 2.078 are positive, 4.309 are negative and 142.223 are neutral. Finally, our fourth baseline belongs to third category and was proposed by Bing Liu (BL) in [6]. It has two lists of words, one positive and other negative with 2.006 and 4.783 words, respectively. It is noteworthy that BL lexicon also presents morphological variants of words and slang. For this analysis, we use the LEGI's resulting lexicon obtained with 100 seeds (the 100 most frequent words in each collection), minimum support 3, confidence 0.8 and maximum distance 3, which corresponds to the best parameter configuration w.r.t. Macro-F1 in our previous evaluations. This configuration resulted in a lexicon with 1,875 positive words, 1,103 negative and 67,674 neutral for the TV Series dataset and 1,966 positive, 2,625 negative and 66,061 neutral words for USA Elections dataset.

The effectiveness of each lexicon is evaluated through traditional classification metrics of quality using the gold standard lexicon as ground truth. More specifically, we derive the $F1$ metric for each class (i.e., positive, negative and neutral), such as presented by Table 4. We observe that both strategies related to the third category (corpus-based methods), BL and LEGI, presented the best results w.r.t. to $F1$ and $Macro\text{-}F1$ for almost all evaluations, with emphasis on those achieved by our strategy. This result demonstrates that taking into account the application context is very important to achieved good lexicons and that our strategy can be effectively instantiated into two distinct domains.

4.4 Evaluation of Lexicon-based Sentiment Analysis

As the primary use of lexicons is on Sentiment Analysis, a deeper analysis of the LEGI's results requires assessing their actual impact on this task. In this sense, we evaluate the LEGI's lexicons, as well the other four baseline lexicons, described in Section 4.3, along with a technique of Sentiment Analysis widely referred in the literature [17]. Assuming the existence of a prior sentiment for each term, given as input, this technique disambiguates the sentiment of terms or expressions in a sentence according to the con-

text in which they are embedded. This is a costly Natural Language Processing that applies a linguistic parser and generates a dependency tree among terms of a sentence to extract a set of features to be used by an AdaBoost classifier. In our analyses, the prior sentiment of each term is given by each evaluated lexicon. Further, such as originally proposed by the authors, whenever a document contains solely sentiment terms from a single class, we classify it as belonging to that class. Otherwise, it is classified as a neutral document. We apply this technique on both datasets and evaluate the quality of the resulting sentiment analysis through Accuracy (i.e., percentage of document correctly classified) and $Macro\text{-}F1$ metrics, considering the gold standard tweet set of each dataset, such as shown by Table 5.

Table 5: Evaluation of lexicon-based sentiment analysis with different lexicons. The best results for each metric are shaded. We observe that LEGI is able to improve lexicon-based sentiment analysis, presenting the best results in all evaluations.

Lexicon source	TV Series		USA Elections	
	Acc.	Mac-F1	Acc.	Mac-F1
Gold Standard	70.75	60.17	53.72	52.19
MPQA	50.18	43.44	39.75	38.90
INQ	28.63	24.44	37.98	24.47
SentiWordNet	58.72	47.34	45.85	42.16
BL	61.32	51.02	50.52	47.48
LEGI	**67.48**	**54.03**	**52.42**	**49.13**

Aiming to provide an upper bound of analysis, Table 5 also presents the results of sentiment analysis when the gold standard lexicon is given as input for the adopted technique. Since there is no gold standard inputs in practical scenarios and that such set, ideally, contains only the actual prior sentiment of each term, this configuration comprises an unfeasible scenario able to provide results near to the optimum for the adopted technique. Comparing the results achieved by using the other lexicons, it is possible to see that the quality of the sentiment analysis obtained by LEGI outperforms the other methods and it is very close to the gold standard lexicon. This result demonstrates that, although simple, our semi-supervised strategy for consolidating context-aware sentiment lexicons is effective, and may improve lexicon-based sentiment analysis techniques.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented LEGI, a corpus-based method for consolidating context-aware sentiment lexicons based on a semi-supervised strategy for propagation of lexicon-semantic classes on a transition graph of terms. Empirical analyses on two distinct domains, derived from Twitter, demonstrate that LEGI outperformed four well-established methods for lexicon consolidation in the literature [16, 12, 6, 1]. Indeed, besides achieving the best approximation to the gold standard lexicon of each domain, in which all terms were manually classified, LEGI presented a small execution time. Further, we found that LEGI's lexicons may improve the sentiment analysis performed by a traditional method in the literature [17]. Such results point out LEGI as a promising method for consolidating lexicons in high demanding scenarios, such as Social Media, wherein scalability and efficiency are of paramount relevance. As future work, we highlight the necessity of improvement on the LEGI's propagation process by considering a broader set of lexicon-semantic classes existing in the literature [16]. Moreover, we intend to investigate more efficient approaches for exploiting lexicons in order to properly infer the sentiment in streams of data.

6. REFERENCES

[1] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of LREC'10*, Valletta, Malta, may 2010.

[2] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proc. of the 12th ACM WWW*, pages 519–528, Hungary, 2003.

[3] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proc. of the 5th LREC*, pages 417–422, 2006.

[4] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proc. of WWW '13*, pages 607–618, Republic and Canton of Geneva, Switzerland, 2013.

[5] D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proc. of the 18th ACM SIGIR Conference*, pages 246–254, Seattle, USA, 1995.

[6] B. Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*, 2010.

[7] S. Mohammad, C. Dunne, and B. J. Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *EMNLP*, pages 599–608. ACL, 2009.

[8] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Sentiful: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 2:22–36, 2011.

[9] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of the AAAI Conference on Weblogs and Social Media*, 2010.

[10] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of LREC'10)*, Valletta, Malta, 2010.

[11] H. Saif, Y. He, and H. Alani. Semantic sentiment analysis of twitter. In *Proc. of 11th ISWC*, pages 508–524, Berlin, Heidelberg, 2012.

[12] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June 2011.

[14] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4), 2003.

[15] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30(3):277– 308, 2004.

[16] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3), 2005.

[17] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of HLT '05*, pages 347–354, USA, 2005.