Debunking Sentiment Lexicons: A Case of Domain-Specific Sentiment Classification for Croatian

Paula Gombar Zoran Medić Domagoj Alagić Jan Šnajder
Text Analysis and Knowledge Engineering Lab
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{paula.gombar,zoran.medic,domagoj.alagic,jan.snajder}@fer.hr

Abstract

Sentiment lexicons are widely used as an intuitive and inexpensive way of tackling sentiment classification, often within a simple lexicon word-counting approach or as part of a supervised model. However, it is an open question whether these approaches can compete with supervised models that use only word-representation features. We address this question in the context of domain-specific sentiment classification for Croatian. We experiment with the graph-based acquisition of sentiment lexicons, analyze their quality, and investigate how effectively they can be used in sentiment classification. Our results indicate that, even with as few as 500 labeled instances, a supervised model substantially outperforms a word-counting model. We also observe that adding lexicon-based features does not significantly improve supervised sentiment classification.

1 Introduction

Sentiment analysis (Pang et al., 2008) aims to recognize both subjectivity and polarity of texts, information that can be beneficial in various applications, including social studies (O'Connor et al., 2010), marketing analyses (He et al., 2013), and stock price prediction (Devitt and Ahmad, 2007). In general, however, building a well-performing sentiment analysis model requires a fair amount of sentiment-labeled data, whose collection is often costly and time-consuming. A popular annotation-light alternative are sentiment polarity lexicons (Taboada et al., 2011): lists of positive and negative words that most likely induce the corresponding sentiment. The key selling points of sentiment lexicons are that they are interpretable and quite easy to be compiled manually. If there is no sentiment-labeled data available, sentiment lexicons can be used directly for sentiment classification: the text is simply classified as positive if it contains more words from a positive than a negative lexicon, and classified as negative otherwise (we refer to this as *lexicon word-counting models*). On the other hand, if sentiment-labeled data is available, sentiment lexicons can be used as (additional) features for supervised sentiment classification models.

One challenge of sentiment analysis is that the task is highly domain dependent (Turney, 2002; Baccianella et al., 2010). This means that generic sentiment lexicons will often not be useful for a specific domain. A notorious example is the word *unpredictable*, which is typically positive in the domain of movie and book reviews, but generally negative in other domains.

The aim of this paper is to investigate how sentiment lexicons work for domain-specific sentiment classification for Croatian. Our main goal is to find out whether sentiment lexicons can be of use for sentiment classification, either as a part of a simple word-counting model or as an addition to a supervised model using word-representation features. To this end, we use a semi-supervised graph-based method to acquire sentiment lexicons from a corpus. We experiment with acquisition parameters, considering both generic and domain-specific seed sets and corpora. We compare all the acquired lexicons with the manually annotated ones. Moreover, we evaluate the lexicon-based models on the task of domain-specific sentiment classification and compare them against supervised models. Finally, we investigate whether a word-counting model can have an edge over a supervised model when the labeled data is lacking.

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 54–59,
Valencia, Spain, 4 April 2017.

2 Related Work

There has been a lot of research on sentiment lexicon acquisition, covering both corpora- and resource-based approaches across many languages (Taboada et al., 2006; Kaji and Kitsuregawa, 2007; Lu et al., 2010; Rao and Ravichandran, 2009; Turney and Littman, 2003). A common approach includes bootstrapping, a method which constructs a sentiment lexicon starting from a small manually-labeled seed set (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003). Moreover, a problem of lexicon domain dependence has also been addressed (Kanayama and Nasukawa, 2006).

Even though most research on sentiment lexicon acquisition and lexicon-based sentiment classification deals with English, there has been some work on Slavic languages as well, including Macedonian (Jovanoski et al., 2015), Croatian (Glavaš et al., 2012b), Slovene (Fišer et al., 2016), and Serbian (Mladenović et al., 2016). While we follow the work of Glavaš et al. (2012b), who focused on the task of semi-supervised lexicon acqusition, we turn our attention to evaluating the so-obtained lexicons on the task of sentiment classification.

3 Lexicon Acquisition

3.1 Dataset

For our experiments, we used a large sentiment-annotated dataset of user posts gathered from the Facebook pages of various Croatian internet and mobile service providers.[1] The dataset comprises 15,718 user posts categorized into three classes: positive (*POS*), negative (*NEG*), and neutral (*NEU*). The average post length is around 25 tokens. We randomly sampled 3,052 posts (245 positive, 1,638 negative, and 1,169 neutral), which we used for lexicon acquisition. The rest of the dataset (12,666 posts) was used for training and evaluation of supervised models.

3.2 Lexicon Construction

We acquired a domain-specific lexicon of unigrams, bigrams, and trigrams (henceforth: n-grams) using a semi-supervised graph-based method. We follow the previous work (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Glavaš et al., 2012b) and employ bootstrapping, which amounts to manually labeling a small set of seed words whose labels are then propagated across the graph. For this, we use a random walk algorithm.

Graph construction. We set all the corpus n-grams as nodes of a graph, which are connected if the words (nodes) co-occur within a same user post in the dataset. For edge weights, we experimented with two strategies: raw co-occurrence counts (co-oc) and pointwise mutual information (PMI). We also filtered out the n-grams that are made solely out of non-content words and that occur less than three times (unigrams) or two times (bigrams and trigrams).

Seed set. We expect that seed selection may affect label propagation in the graph. To investigate this, we experimented with different seed sets, each containing 3×15 n-grams (15 n-grams per class):

- Two generic, human-compiled seed sets (GH1, GH2) – Two Croatian native speakers compiled the generic seed sets following their intuition;
- Two domain-specific, human-compiled seed sets (DH1, DH2) – Two Croatian native speakers compiled the seed sets from frequency-sorted list of n-grams from the domain corpus following their intuition;
- One domain-specific, corpus-based seed set (DC1) – Starting from the 45 most frequent n-grams, we circularly assigned one n-gram to the positive, negative, and the neutral seed set, until all n-grams were exhausted (a *round-robin* approach). We used this seed set as a baseline.

An example of a domain-specific, human-compiled seed set is shown in Table 1.

Sentiment propagation. To propagate sentiment labels across graph nodes, we used the PageRank algorithm (Page et al., 1999). Since PageRank was originally designed to rank web pages by their relevance, we adapted it for sentiment propagation, following (Esuli and Sebastiani, 2007; Glavaš et al., 2012a). In each iteration, node scores were computed using the power iteration method:

$$\mathbf{a^{(k)}} = \alpha \mathbf{a^{(k-1)}} \mathbf{W} + (\mathbf{1} - \alpha)\mathbf{e}$$

where $\mathbf{W}$ is the weighted adjacency matrix of the graph, $\mathbf{a}$ is the computed vector of node scores, $\mathbf{e}$

[1] At this point, this dataset is not publicly available as it was constructed within a commercial project. The dataset may be open-sourced in the future.

	Croatian	Translation
Positive seeds	*hvala, zanimati, nov, dobar, brzina, super, lijepo, zadovoljan, besplatan, ostati, riješiti, biti zadovoljan, uredno, brzi, hvala vi*	*thanks, to interest, new, good, speed, super, nice, satisfied, free, to stay, to solve, to be satisfied, tidy, fast, thank you*
Negative seeds	*nemati, problem, ne moći, kvar, ne raditi, čekati, biti problem, prigovor, raskid, katastrofa, sramota, zlo, raskid ugovor, otići, smetnja*	*to not have, problem, to not be able, malfunction, to not work, to wait, to be a problem, objection, break-up, catastrophe, shame, evil, contract termination, to leave, nuisance*
Neutral seeds	*imati, dan, internet, broj, korisnik, mobitel, ugovor, tarifa, mjesec, poruka, nov, vip, reći, poziv, signal*	*to have, day, internet, number, user, cellphone, contract, rate, month, message, new, vip, to say, call, signal*

Table 1: Human-generated domain-specific seed set (lemmatized).

is a vector of normalized internal node scores, and α is the damping factor (we used a default value of 0.15). In the initialization phase, the adjacency matrix $\mathbf{W}$ was row-normalized and the nodes from the seed set were set to $\frac{1}{|SeedSet|}$, whereas the rest of the nodes were set to 0.

We then ran the algorithm twice, once with positive seeds and once with negative ones, obtaining ranked lists of positive and negative scores of all n-grams. To determine the final sentiment of an n-gram, we first calculated the difference between its ranks in the lists of positive and negative scores, and then compared it to a fixed threshold. If the difference between its ranks was below the threshold, the n-gram was classified as neutral. If not, it was classified as positive if its rank was higher in the list of positive scores and negative otherwise. We also tried using score difference, but rank difference worked better. Lastly, it is worth noting that, as the goal of our work is to determine the best possible performance of a lexicon-based sentiment classifier, we computed an oracle threshold by optimizing the threshold on the gold set, as described in the following section.

3.3 Lexicon Evaluation

Gold lexicon construction. We made use of our sentiment-labeled dataset to extract the most representative subset of n-grams for the annotation. More precisely, we ranked all the n-grams according to their χ^2 scores, which were calculated based on their co-occurrence with *POS*, *NEU*, and *NEG* user posts in the dataset. To obtain a final list of n-grams for the annotation, we selected 1,000 n-grams by uniformly sampling all these three lists from the top, making sure to avoid duplicates. Subsequently, five annotators labeled the dataset, and we obtained the final label as a majority vote (there were no ties).

Parameter	Optimal value
Weighting strategy	Raw co-occurrence counts
Seed set	DH2
Classification strategy	Rank difference
Classification threshold	77

Table 2: Parameters used for obtaining the best-performing domain-specific lexicon when evaluated against the gold lexicon.

	Generic		Domain-specific		
	GH1	GH2	DH1	DH2	DC1
Co-oc	37.9	40.0	43.8	**46.2**	38.3
PMI	36.7	38.1	39.9	45.0	35.8

Table 3: F1-scores of acquired lexicons evaluated against the gold lexicon.

Inter-annotator agreement. We measured the inter-annotator agreement (IAA) using both the Cohen's kappa (Cohen, 1960) and pairwise F1-score. We first calculated the agreement for all annotator pairs and averaged them to obtain the overall agreement. The averaged Cohen's kappa is 0.68, which is considered a substantial agreement, according to Landis and Koch (1977). The macro-averaged F1-score is 0.79.

Evaluating generated lexicons. We have acquired a total of 10 lexicons, combining two weighting strategies (raw co-occurrence count and PMI) with five different seed sets (cf. Section 3.2). We evaluated these against the human-annotated gold lexicon in terms of macro-averaged F1-score. Using optimal parameters from Table 2, we obtained the score of 0.46. The other lexicons' scores are reported in Table 3.

Seed-corpus type	P	R	F1
domain-domain	42.1	41.66	39.79
generic-domain	45.31	46.01	**44.77**
generic-generic	17.39	33.33	22.85

Table 4: Scores of word-counting models.

4 Sentiment Classification

After obtaining the optimal lexicon (in comparison to the gold lexicon), we test how well it performs on the task of sentiment classification of user posts. This task commonly incorporates sentiment lexicons in two ways: as a part of a simple word-counting approach, or as a source of lexicon-based features in a supervised model. We are interested in how simple word-counting approach fares against the more complex supervised one. The models are evaluated using a nested k-fold cross-validation (10×5 folds) on the subset of our sentiment-labeled dataset that was not used for lexicon acquisition.

4.1 Lexicon Word-Counting Classification

In this setup, a user post is classified as positive if it contains more positive than negative n-grams from the lexicon, and vice versa. In case of ties, the user post is predicted neutral. To investigate how different seed sets and corpora influence lexicon quality, we compare our best-performing lexicon (*domain-domain*;[2] Co-oc DH2) to two additional lexicons: a domain-specific lexicon built with generic seeds (*generic-domain*; Co-oc GH2) and a generic Croatian lexicon compiled by Glavaš et al. (2012b) (*generic-generic*).

We evaluated the models in terms of macro-averaged F1-scores, which we report in Table 4. Surprisingly, the *generic-domain* lexicon outperformed the one that seemed the best when compared against the gold lexicon (*domain-domain*).

4.2 Supervised Classification

For the supervised classification, we decided to use a simple logistic regression model with lexicon-based and word-representation features. Lexicon-based features capture how many words from the positive and negative lexicon appeared in a user post, as well as the average rank and score of words from the positive and negative lexicons. On the other hand, for word-representation features we use tf-idf-weighted bag-of-words vectors (BoW) and the popular skip-gram embeddings (SG) proposed by Mikolov et al. (2013). We build 300-dimensional vectors from hrWaC, a Croatian web corpus (Ljubešić and Erjavec, 2011), filtered by Šnajder et al. (2013) using the `word2vec` tool.[3] We set the negative sampling parameter to 5, minimum frequency threshold to 100, and we did not use hierarchical softmax. To construct user post skip-gram embeddings, we follow the common practice and average the embeddings of its content words.

Model	P	R	F1
domain-domain	63.82	43.01	**41.98**
generic-domain	39.19	41.11	39.08
SG	64.57	58.20	60.27
SG + generic-domain	65.60	59.39	61.42
SG + domain-domain	65.70	59.48	**61.53**
BoW	69.93	63.55	65.75
BoW + generic-domain	70.08	63.22	65.50
BoW + domain-domain	70.68	63.47	**65.90**

Table 5: Scores of supervised models with lexicon-based and word-representation features.

For the evaluation, we decided to omit the *generic-generic* lexicon from our experiments due to its subpar performance in lexicon word-counting classification. To see how lexicon-based features affect the classification performance, we evaluate models that use them in conjunction with word-representation features and models that use them as the only features. The boost in the models' scores when using both types of features is not statistically significant (paired t-test with $p<0.001$). We report the scores in Table 5.

4.3 Discussion

Based on the results from Tables 4 and 5, we observe that any supervised model based on word-representation features (with or without lexicon-based features) greatly outperforms word-counting models and models based on lexicon-based features. This indicates that, in our case, it makes sense to use a simple word-counting model (F1-score of 44.77%) when annotating data is entirely infeasible, and a supervised model with word-representation features in all other cases (F1-score of 65.90%).

It is interesting to investigate whether the above

[2] Here, *domain-domain* refers to a lexicon built with a domain-specific seed set over a domain-specific corpus.

[3] `https://code.google.com/archive/p/word2vec/`

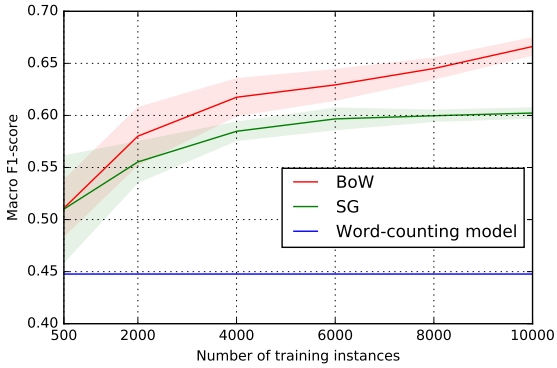

Figure 1: Learning curves of the supervised models (BoW and SG) and the word-counting model.

observation holds even when dealing with a relatively small amount of sentiment-labeled data. To that end, we inspect the learning curve of these models' performances (Figure 1). We observe that annotating as few as 500 instances already makes both supervised models outperform the lexicon word-counting model by a large margin.

5 Conclusion

We tackled the domain-specific sentiment lexicon acquisition and sentiment classification for Croatian. We used a semi-supervised graph-based model to acquire lexicons using both generic and domain-specific seed sets and corpora. Furthermore, we analyzed their quality against the human-annotated gold lexicons. Within the context of domain-specific sentiment classification, we used the obtained lexicons both as part of a lexicon word-counting model and as features for a supervised model, and showed that they do not yield any significant improvements. Finally, we reported that, even in the case of having as few as 500 labeled instances, simple word-counting models cannot compete with supervised models based on word-representation features. For future work, we plan to carry out a more extensive analysis across several different domains and languages.

Acknowledgments

The research has been carried out within the project "CATACX: Cog-Affective social media Text Analytics for Customer eXperience analysis (PoC6-1-147)", funded by the Croatian Agency for SMEs, Innovations and Investments (HAMAG-BICRO) from the Proof of Concept Program.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2200–2204, Valletta, Malta.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 984–991, Prague, Czech Republic.

Andrea Esuli and Fabrizio Sebastiani. 2007. Pageranking wordnet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, volume 7, pages 442–431, Prague, Czech Republic.

Darja Fišer, Jasmina Smailović, Tomaž Erjavec, Igor Mozetič, and Miha Grčar. 2016. Sentiment annotation of Slovene user-generated content. In *Proceedings of the 2016 Conference Language Technologies and Digital Humanities (JTDH 2016)*, pages 65–70, Ljubljana, Slovenia.

Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić. 2012a. Experiments on hybrid corpus-based sentiment lexicon acquisition. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 1–9, Avignon, France.

Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić. 2012b. Semi-supervised acquisition of Croatian sentiment lexicon. In *International Conference on Text, Speech and Dialogue*, pages 166–173. Springer.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics (EACL 1997)*, pages 174–181, Madrid, Spain.

Wu He, Shenghua Zha, and Ling Li. 2013. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472.

Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2015. Sentiment analysis in Twitter for Macedonian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2015)*, pages 249–257, Hissar, Bulgaria.

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 11th Conference on Computational Natural Language Learning (CoNLL 2007)*, pages 1075–1083, Prague, Czech Republic.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 355–363, Sydney, Australia.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In *Proceedings of 14th International Conference on Text, Speech and Dialogue (TSD 2011)*, pages 395–402, Pilsen, Czech Republic.

Bin Lu, Yan Song, Xing Zhang, and Benjamin K Tsou. 2010. Learning Chinese polarity lexicons by integration of graph models and morphological features. In *Asia Information Retrieval Symposium*, pages 466–477. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Neural Information Processing Systems Conference (NIPS 2013)*, pages 3111–3119, Lake Tahoe, USA.

Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2016. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: bringing order to the web.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 675–682, Athens, Greece.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for Croatian. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Maite Taboada, Caroline Anthony, and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 427–432, Genoa, Italy.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 417–424, Philadephia, Pennsylvania, USA.