Informal sentiment analysis in multiple domains for English and Spanish

Tadej Štajner[1,2]*, Inna Novalija*[1]*, Dunja Mladenić*[1,2]

[1]Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773900
e-mail: {firstname.secondname}@ijs.si

[2]Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773100

ABSTRACT

This paper addresses the problem of sentiment analysis in an informal setting in multiple domains and in two languages. We explore the influence of using background knowledge in the form of different sentiment lexicons, as well as the influence of various lexical surface features. We show that the improvement resulting from using a two-layer model, sentiment lexicons, surface features and feature scaling is most notable on social media datasets in both English and Spanish. For English, we are also able to demonstrate improvement on the news domain using sentiment lexicons and a large improvement on the social media domain. We also demonstrate that domain-specific lexicons bring comparable performance to general-purpose lexicons.

1 INTRODUCTION

Sentiment analysis is a natural language processing task which aims to predict the polarity (positive, negative or neutral) of users publishing sentiment data, in which they express their opinions. The task is traditionally tackled as a classification problem using supervised machine learning techniques. However, this approach requires additional effort in manual labeling of examples and often has difficulties in transferring to other domains.

One way to ameliorate this problem is to construct a lexicon of sentiment-bearing words, constructed from a wide variety of domains. While some sentiment-bearing cues are contextual, having different polarities in different contexts, the majority of words have unambiguous polarity. While this is a compromise, research shows that lexicon-based approaches can be an adequate solution if no training data is available. In practice, sentiment dictionaries or lexicons are lexical resources, which contain word associations with particular sentiment scores. Dictionaries are frequently used for sentiment analysis, since they allow in a fast and effective way to detect an opinion represented in text. While there exists a number of sentiment lexicons in English [1][2], the representation of sentiment resources in other lexicons is not as developed.

The second problem this paper focuses on is detecting sentiment in social media. Besides being domain-specific, it can also be grammatically less correct and contain other properties, such as mentions of other people hash-tags, smileys and URL, as opposed to traditional movie and product review datasets.

This paper explores various combinations of methods that can be used to incorporate out-of-domain training data, combined with lexicons in order to train a domain-specific sentiment classifier.

2 RELATED WORK

Sentiment classification is an important part of our information gathering behavior, giving us the answer to what other people think about a particular topic. It is also one of the natural language processing tasks which is well suited for machine learning, since it can be represented as a three-class classification problem (positive, neutral, negative). Earlier work applied sentiment classification to movie reviews [10], training a model for predicting whether a particular review rates a movie positively or negatively. While in the review domain all examples are inherently either positive or negative, other domains may also deal with non-subjective content which does not carry any sentiment. Furthermore, separating subjective from objective examples has proven to be an even more difficult problem than separating positive from negative examples [13]. Another difficult problem in this area is dealing with different topics and domains: models, trained on a particular domain do not always transfer well onto other domains. While the standard approach is to use one of widely used classification algorithms such as multinomial Naïve Bayes or SVM, explicit knowledge transfer approaches have been proven to improve performance in these scenarios, such as using sentiment lexicons [1] or modifying the learning algorithm to incorporate background knowledge [9]. Some challenges are also domain-specific. For instance, while a lot of sentiment is being expressed in social media, the language is often very informal, affecting the performance by increasing the sparsity of the feature space. On the other hand, the patterns arising in informal communication, such as misspellings and emoticons can be themselves used as signals [13]. It has also been shown that within social media, using different document sources, such as blogs, microblogs and reviews, can improve performance compared to using a single source. [12]

3 SENTIMENT LEXICONS

SentiWordNet [1] is the most known English-language sentiment dictionary, in which each WordNet [3] synset s is represented with three numerical scores – objective Obj(s), positive Pos(s) and negative Neg(s). However, SentiWordNet does not account for domain specificity of the input textual resources. In addition to addressing English language, this paper also discusses applications of sentiment dictionaries in Spanish. For this purpose, we have used the sentiment dictionaries published by Perez-Rosas et al. [6].

Expressing sentiment and opinion varies for different domains and document types. In such way, sentiments carried in the news are not equivalent to the sentiments from the Twitter comments. For instance, the word “turtle” is neutral in a zoological text, but in informal Twitter comment “connection slow as a turtle”, “turtle” has negative sentiment. This paper also evaluates a method for construction of dictionaries as domain specific lexical resources, which contain words, part of speech tags and the relevant sentiment scores. We have set the topic of telecommunications as the domain of primary interest, and the corpus, used for dictionaries development, was composed out of Twitter comments about telecommunication companies. We have started with a number of positive and negative seeds for different part-of-speech words (adjectives, nouns, verbs). These sentiment dictionaries are built in English and Spanish languages. As discussed in [3], there are a number of approaches to develop the sentiment dictionary. In our research on developing sentiment dictionaries we were following the work of Bizau et al. [4]. In the paper on expressing opinion diversity, the authors suggested a 4-step methodology for creating a domain specific sentiment lexicon. We have modified the methodology in order to generalize to other languages and provide sentiments for different parts of speech.

We have created dictionaries not only in English, but also in Spanish. Our dictionaries were built not only for adjectives as done in [4], but also for nous and verbs. For the English dictionary, we have additionally provided several extra features, such as the number of positive links and number of negative links for a particular word. The English sentiment dictionary for the Telecommunication domain is composed out of around 2000 adjectives, 1700 verbs and 8000 nouns, while the Spanish counterpart contains around 650 adjectives, 2000 verbs and 4100 nouns.

4 FEATURE CONSTRUCTION

We have used different feature sources to represent individual opinion data points. In news and review datasets, every data point is a sentence, while in social media datasets, every data point is a single microblog post. We preprocess the textual contents by replacing URLs, numerical expressions and the names of opinions’ targets with respective placeholders. We then tokenize this text, lower-casing and normalizing characters onto an ASCII representation, filtering for stopwords and weigh the terms using TF-IDF weights. The words were stemmed using the Snowball stemmer for English and Spanish. The punctuation is preserved.

To accommodate social media, we have also used other text-derived features that can carry sentiment signal in informal settings:

- count of fully capitalized words
- count of question-indicating words
- count of words that start with a capital letter
- count of repeated exclamation marks
- count of repeated same vowel
- count of repeated same character
- proportion of capital letters
- proportion of vowels
- count of negation words
- count of contrast words
- count of positive emoticons
- count of negative emoticons
- count of punctuation
- count of profanity words[1]

We use lexicons in the form of features, where every word has assigned one or more scores. For instance, our dictionaries, described in Section 3, as well as SenticNet, provide a single real value in the range from -1 to 1, representing the scale from negative to positive. For these lexicons, we generate the sum of sentiment scores and the sum of absolute values of sentiment scores for every part of speech tag, as well as in total. SentiWordNet scores are represented as a triple of positive, negative and objective scores, having a total sum of 1.0. We have used a similar feature construction process as in [7]: providing sums of positive and negative scores, as well as the ratio of positive to negative score. These features were computed for each part of speech tag and in total. For Spanish, we have used the UNT sentiment lexicon [6]. Since each entry is labeled as positive or negative, we use the count of detected positive words and count of detected negative words as features.

5 MODELS

The data is composed of two modalities: bag-of-words features on one side, and having lexical and surface features, such as patterns and lexicon features on the other. In order to take differing distributions into account, we use two different approaches: either concatenating the features into a single features space, or using different models for each set of features. While this situation has been solved by extending the Naïve Bayes classifier with pooling multinomials [9], we chose to implement it with a two-step model. While they demonstrate that Multinomial Naïve Bayes performs well in sentiment analysis tasks, our results show that combining bag-of-words with lexical and surface feature reduces performance instead of improving it. We therefore experiment with modeling approaches that are better suited for integration of background knowledge.

[1] Obtained from http://svn.navi.cx/misc/abandoned/opencombat/misc/multilingualSwearList.txt

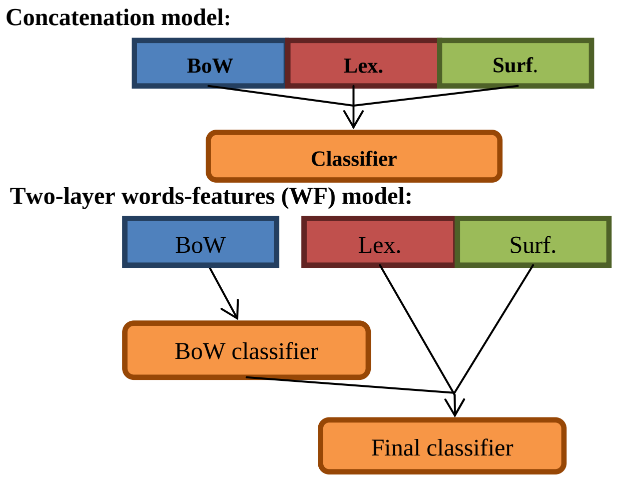

Figure 1: *Diagrams of the simple concatenation model and the two-layer words-features model which encodes the BoW model output as features for the final model.*

We therefore compare two modeling approaches, illustrated in Figure 1. We experiment by varying the training algorithm used: for the concatenating model, we vary the main algorithm, and for the two-layer model, we vary the second level algorithm, as we have fixed the BoW level classifier to Linear SVM, known to work well on BoW.

6 EXPERIMENTS

Furthermore, we focus our experiment onto performance on our target datasets. We use the following datasets:

- Pang & Lee review dataset, English [10]
- JRC news dataset, English [11]
- JRC news dataset, translated to Spanish using Microsoft Translator (JRC-ES)
- RenderEN, English. 134 Twitter posts about a telecommunications provider (48 Pos, 84 Neg)
- RenderES, Spanish, 891 Twitter posts about a telecommunications provider (388 Pos, 445 Neg, 58 Obj)

Besides our lexicons introduced in section 3 (denoted “RenLex” and “RenLexLinks”), we also evaluate performance of using the Spanish lexicons from Perez-Rosas et al [6] (denoted FullUNT and MedUNT for the full and medium variant respectively), as well as SenticNet [8] and SentiWordNet[1] for English. The label “Lex” indicates usage of all lexicons. Our key indicators are performance metrics on RenderEN and RenderES, as they represent our use case. We report F_1 scores for all of these datasets on various combinations of classifiers and features construction schemes. The experiments cover various learning algorithms, both modeling pipelines (“WF-“ denotes the two-layer model), as well as the effect of feature scaling and centering (denoted with “WF-SVMSc”). We explore various combinations of feature sets: surface, bag-of-words, lexicons, as well as performance of individual lexicons.

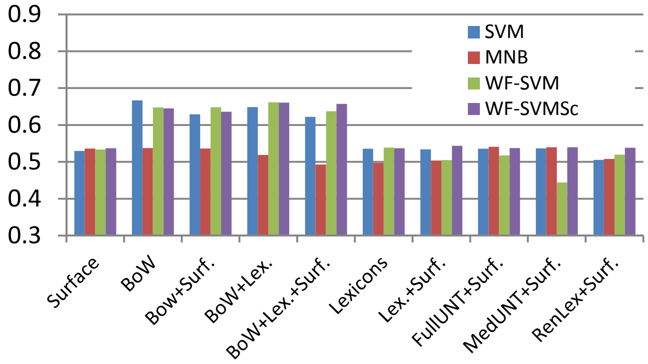

Table 1: Sentiment F_1 *scores on JRC-ES across settings.*

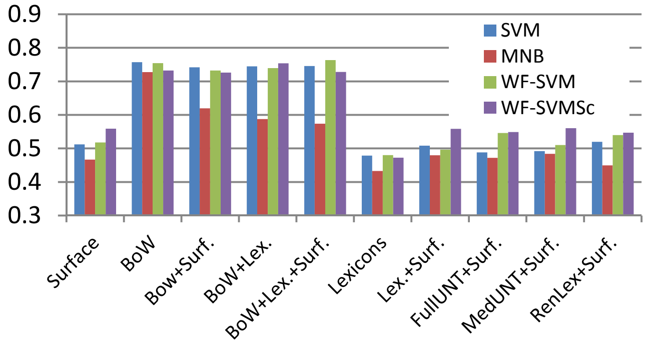

Table 2: Sentiment F_1 *scores on Render-ES across settings.*

Table 1 and 2 present the results on both Spanish datasets when combining different feature sets and learning approaches. We observe that on the news dataset, none of the additions improve over the bag-of-words baseline on an SVM model at 0.66 F_1 score. On Render-ES, the variant combining all additions and running on a two-layer SVM model improves over the bag-of-words model by a small margin, resulting in an F_1 score of 0.76. Looking at usage of various lexicons alone, it shows that the lexicons themselves only slightly improve over the surface features. In many cases, the difference is not significant, although we observe that the domain specific lexicon RenLex does not improve over a general domain lexicon neither in news nor in social media.

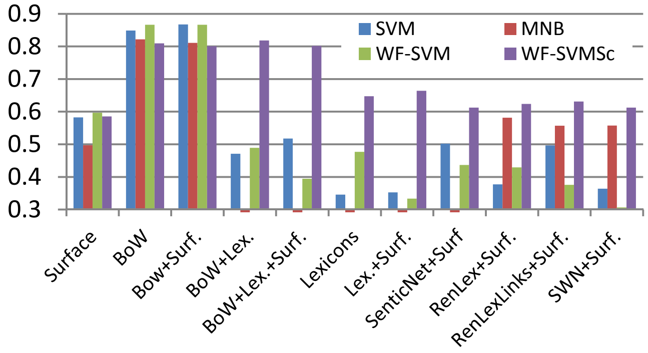

Table 3: Sentiment F_1 *score on PangLee across settings.*

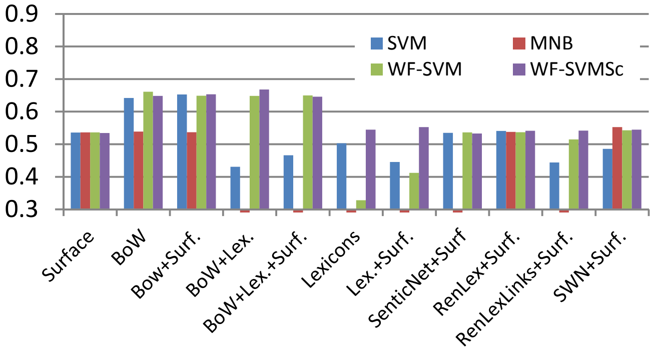

Table 4: Sentiment *F_1 scores on JRC-EN across settings.*

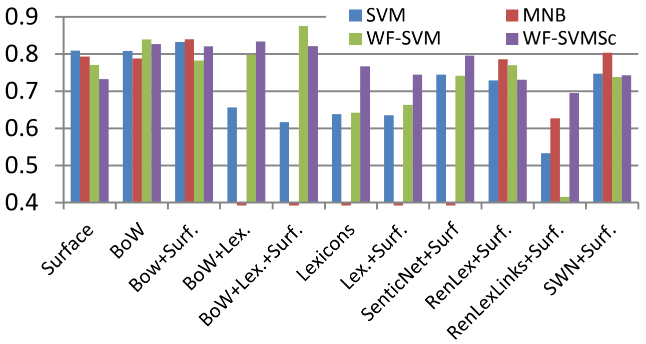

Table 5: Sentiment *F_1 scores on Render-EN across settings.*

Tables 3, 4 and 5 show the results on English reviews, news, and social media. While none of the additions beat the bag-of-words baselines on reviews, scoring at 0.86, it demonstrates that when combining bag of words and lexicon features, the two-step WF model is more robust than concatenation. It also demonstrates the importance of feature centering when combining lexicon features with outputs from the bag-of-words model. On news, while adding lexicons improves the performance from 0.66 to 0.67, surface features don't give any improvement, mostly due to the formal language used in reporting. On the final, social media dataset, we demonstrate the performance improvements in combining all three feature sets in a two-layer model along with feature scaling. The best performing model is able to obtain a F_1 score of 0.88. While the dataset is small, this demonstrates the feasibility of using external knowledge and surface features in a social media setting, especially with insufficient training data. Also, using the number of positive and negative links as features does not improve performance.

7 CONCLUSIONS

Results confirm that social media content is the domain which benefits the most from external knowledge. We show that topic-specific lexicons don't bring improvement over general purpose lexicons, likely because the ambiguity of certain words that a topic-specific lexicon would solve was not problematic. We have been able to show improvement on two English datasets, especially on social media, which benefited significantly from preprocessing, surface features, as well as lexicons. We also demonstrate feasibility of using machine translation to obtain a training corpus in another language. Evaluation shows that the performance for JRC-ES was comparable to JRC-EN. Other research shows [9] promising approaches to facilitate the knowledge transfer via lexicons using specifically tailored machine learning approaches. In future work we will explore cross-lingual learning, demonstrating approaches for training sentiment models using language resources from other languages.

7 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (ICT-216886-NoE), XLike (ICT-STREP-288342), and RENDER (ICT-257790-STREP).

References

[1] Esuli, A. and Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th LREC.

[2] Janyce Wiebe and Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Proceeding of CICLing-05, pages 486–497, Mexico City, Mexico.

[3] Fellbaum, Ch. 1998. WordNet: An Electronic Lexical Database. MIT Press.

[4] Bizau, A., Rusu, D., Mladenic. D. 2011. Expressing Opinion Diversity. In Proceedings of the 1st Intl. Workshop on Knowledge Diversity on the Web (DiversiWeb 2011), Hyderabad, India.

[5] Hatzivassiloglou, V. and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the ACL.

[6] Perez-Rosas, V., Banea, C., Mihalcea, R: Learning Sentiment Lexicons in Spanish. In Proceedings of the LREC 2012

[7] Ohana, B. and Tierney, B: Sentiment classification of reviews using SentiWordNet, In Proceedings of 9th. IT & T Conference, 2009

[8] E. Cambria, C. Havasi, and A. Hussain. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In: Proceedings of FLAIRS, pp. 202-207, Marco Island (2012)

[9] Melville, P. and Gryc, W. and Lawrence, R.D.: Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. Proceedings of the 15th ACM SIGKDD, 2009

[10] Pang, B., Lee, L., and Vaithyanathan, S: Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.

[11] Balahur, A. and Steinberger, R. and Kabadjov, M. and Zavarella, V. and Van Der Goot, E. and Halkia, M. and Pouliquen, B. and Belyaeva, J:. Sentiment Analysis In the News. Proceedings of LREC, 2010

[12] Yelena Mejova, Padmini Srinivasan: Crossing Media Streams with Sentiment: Domain Adaptation in Blogs, Reviews and Twitter. In Proceedings of the 6th ICWSM, ACM, 2012

[13] Bo Pang, Lillian Lee: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.