Informal Multilingual Multi-domain Sentiment Analysis

Tadej Štajner[1,2], Inna Novalija[1] and 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

Keywords: sentiment analysis, social media, news sentiment, opinion mining

Received: March 6, 2013

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 evaluate several different feature set combination strategies. We show that the improvement resulting from using a two-layer meta-model over the bag-of-words, sentiment lexicons and surface features 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 as well as a large improvement on the social media domain. We also demonstrate that domain-specific lexicons bring comparable performance to general-purpose lexicons.

Povzetek: Ta članek obravnava problem analize naklonjenosti v neformalnem besedilu v različnih domenah in v dveh različnih jezikih.

1 Introduction

Sentiment analysis is a natural language processing task which aims to predict the polarity (usually denoted as 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 labelling 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 languages is not as developed. The first problem this paper focuses on is integrating external knowledge in the form of general-purpose sentiment lexicons.

The second problem this paper focuses on is detecting sentiment in specific domains, such as 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 behaviour, 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, classifying every example into either positive, neutral, or 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] .

This paper also explores the integration of multiple data representations for a specific task of text classification. This sort of approach was also successful in the case where several combination strategies were used for the task of authorship detection [14] , such as feature set concatenation or majority voting of classifiers, trained on only subsets of features. While these are known general strategies, a lot of aspects of selecting sensible feature subsets are very domain specific.

3 Sentiment Lexicons

SentiWordNet [1] is the most known English-language sentiment dictionary, in which each WordNet [3] synset 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 chosen the topic of telecommunication services within social media as the domain of primary interest, and the corpus, used for dictionaries development, was composed out of Twitter comments referring to services of 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], where, 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 [17] . The punctuation is preserved.

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

- 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]

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

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] :

- Sum of all positive sentiments of all words.
- Sum of all negative sentiment of all words.
- Total objective sentiment of all words (where *obj = 1.0 - (pos + neg)*) score
- Ratio of total positive to negative scores for all words

Besides providing total sums, we also generate these features for nouns, verbs, adjectives and adverbs separately.

For Spanish, we have used the UNT sentiment lexicon [6] . Since each entry is labelled only as either 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 three main modalities: bag-of-words features, lexicon features, and surface features.. In order to take differing distributions, dimensionality and sparsity properties 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. We experiment with different feature combination approaches that are better suited for integration of background knowledge and other learning algorithms.

5.1 Feature combination

We therefore compare three feature combination approaches and a baseline, illustrated in Figures 1 through 4. The concatenating model simply stacks all feature spaces together and performs learning on the joint feature space. While this approach is simple, it is sensitive to different feature distributions. Therefore, we pre-emptively scale the features, so every feature has a standard deviation of 1.0. We don't standardize the mean, since the features themselves may be sparse, and complete standardization would densify the data. The concatenation approach from Figure 1 is considered as the baseline.

The second approach, as shown in Figure 2, is using a separate learning model for the bag-of-words feature set, and feeding the output of that model as features into the final classifier, together with the less sparse lexicon and surface features, in order to 'compress' the bag-of-words signal.

The third approach is related to the well-known attribute bagging [16] meta-learning strategy, with the crucial difference that the feature sub-sets are already defined in advance via domain knowledge.

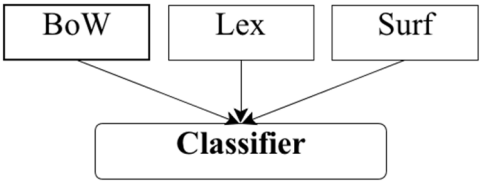

Figure 1: Feature concatenation diagram.

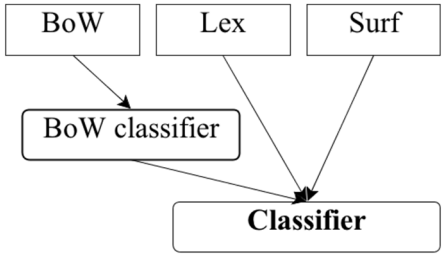

Figure 2: Separate bag-of-words model, denoted as "Words and features"

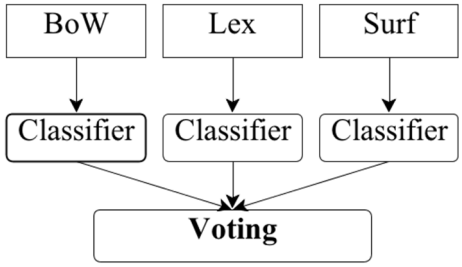

Figure 3: Separate model for every feature set, aggregated by voting.

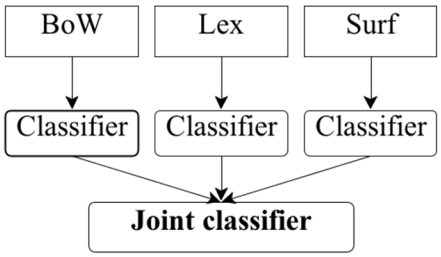

Figure 4: Meta-classifier, using class probabilities from the inner classifier predictions as its features.

The fourth approach extends the voting by employing a separate classifier model that operates on the output of the output probabilities of the inner models, in order to minimize bias of individual feature sets.

We experiment by varying the training algorithm used: For the approaches using multiple models, we use the same algorithm for all the models.

All in all, we evaluate four feature set combination strategies, corresponding to Figures 1-4:

- Concatenation (**Concat**)

- Two-layer words and features (**W+F**)
- Voting model (**Voting**)
- Meta-classifier (**Meta**)

6 Experiments

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

- Pang & Lee review dataset (PangLee), English [10] , consisting of movie reviews, gathered from IMDB.
- JRC news dataset (JRC-en), English [11] , consisting of statements from news articles on the topic of global politics.
- JRC news dataset, translated to Spanish using Microsoft Translator (JRC-es)
- RenderEN, English. 134 Twitter posts about a telecommunications provider (48 positive, 84 negative)
- RenderES, Spanish, 891 Twitter posts about a telecommunications provider (388 positive, 445 negative, 58 objective)

Besides our lexicons introduced in section 3 (denoted "RenderLex" and "RenderLexLinks"), 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 perform experimental evaluation for all of these datasets on various combinations of classifiers and features construction schemes. The experiments cover various learning algorithms, as well as different modelling pipelines. We explore various combinations of feature sets: surface, bag-of-words, lexicons, as well as performance contributions of individual lexicons.

The first evaluation deals with observing the applicability of various sentiment lexicons, as described in Section 3. First, we evaluate the lexicons in isolation, followed by a combination of lexicons together with surface features. We train a L1-regularized logistic regression classifier on lexicon features. The performance is measures using averaged F_1-score [18] in a 10-fold cross-validation setting.

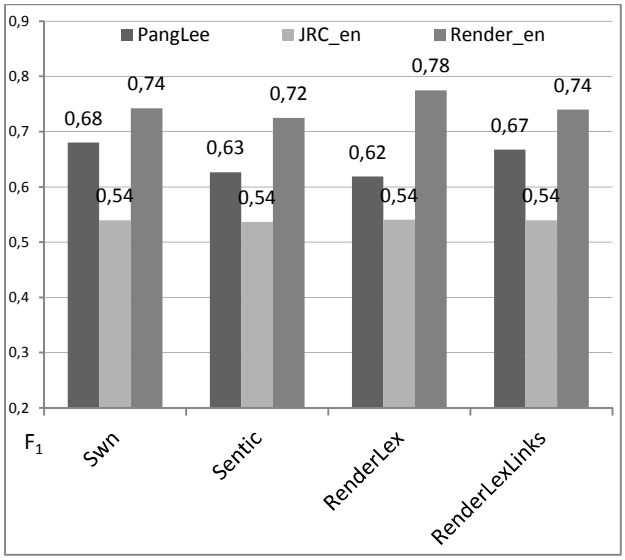

Figure 5: Sentiment F_1 scores with various sentiment lexicons for English.

Figure 5 shows the results, obtained performing sentiment classification on the basis of sentiment lexicon features alone. We observe that performance across the news dataset is constant, since the expression of sentiment in news doesn't directly correspond to sentiment meaning of individual words, but more to the domain-specific political statements. For the social media dataset, we observe improved performance when using a telecommunications domain-specific lexicon, compared to using a general domain sentiment lexicon.

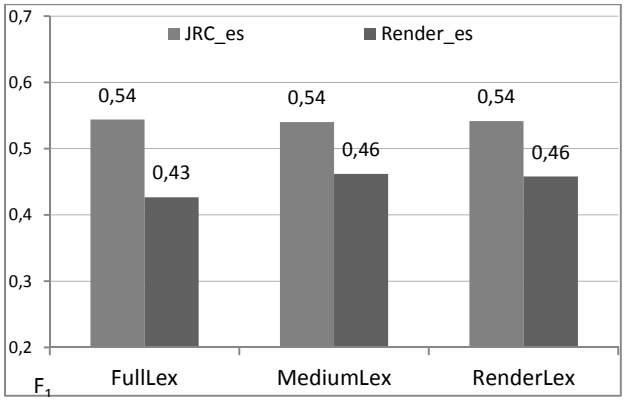

Figure 6: Sentiment F_1 scores with various sentiment lexicons for Spanish.

While Figure 6 confirms the same behaviour for news, the benefit of using lexicons is much lower in Spanish social media content. Given these results, we establish that a custom-built lexicon can give better results than a general purpose one. To continue, we evaluate various feature combination techniques on different learning algorithms.

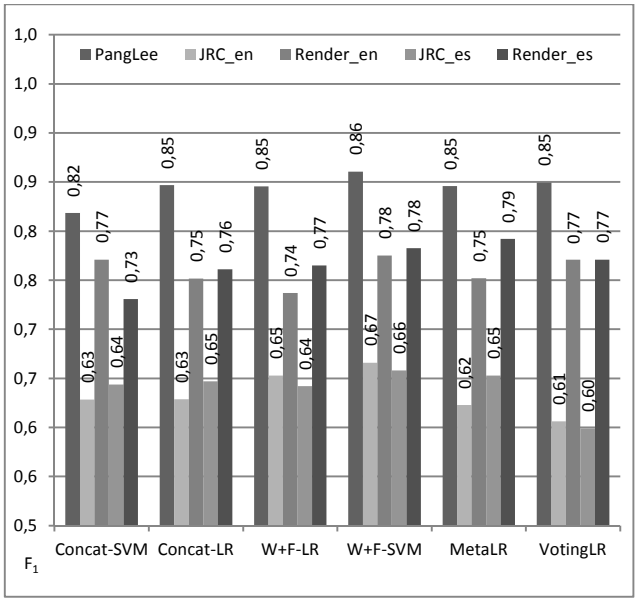

Figure 7: F_1 scores with various feature combination approaches across both languages and two learning approaches.

Figure 7 displays the performance across different feature combination approaches across all datasets. Looking into individual models, we observe that the W+F model, having the bag-of-words feature set on a separate layer, consistently works best for the purpose of combining all the three feature sets and masking the differences in the distribution of their features. While the W+F model consistently outperforms concatenation by a small but statistically significant margin, the Voting or Meta-classifier model only outperform concatenation on some occasions, and perform worse on the news dataset in both languages. We report the results on scenarios where LR was used as the learning algorithm on the Meta and Voting models due to the fact that they obtain comparable performance.

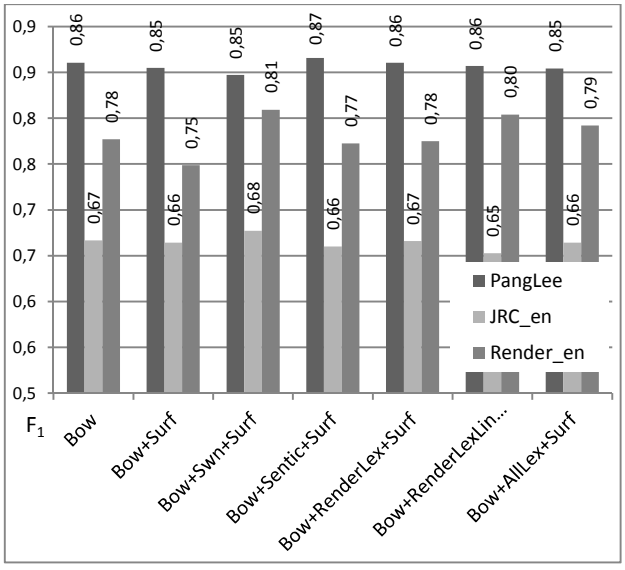

Figure 8: Using various feature sets on English datasets, using W+F-SVM.

Figure 8 shows the results on English reviews, news, and social media. On reviews, none of the additions significantly beat the bag-of-words baselines on reviews. On news, while adding SentiWordNet marginally improves the performance from 0.67 to 0.68, surface features don't give any improvement, mostly due to the formal language used in reporting, which leads to the fact that the text is written without informal cues. On other hand, results on the *Render_en* social media dataset, demonstrate the performance improvements in combining all three feature sets in a two-layer model. The best performing model is able to obtain a F_1 score of 0.87. While the dataset is small, this demonstrates the feasibility of using generalized external knowledge and surface features in a social media setting, especially with insufficient training data.

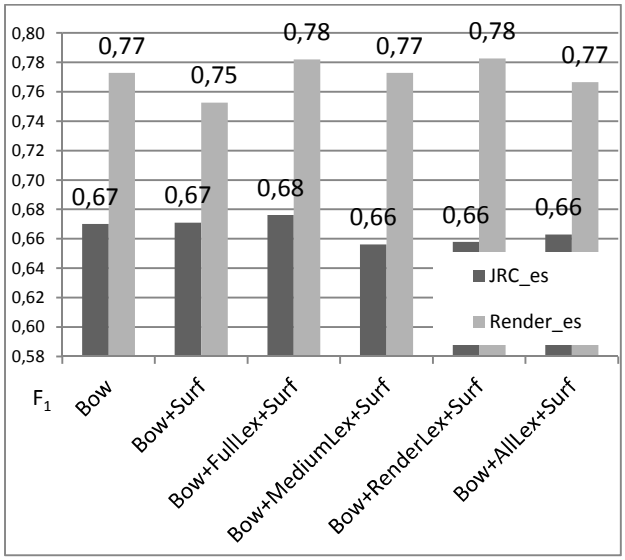

Figure 9: Sentiment F_1 scores on Spanish datasets, using W+F-SVM.

Figure 9 shows the results on both Spanish datasets when combining different feature sets in a W+F setting and a SVM model. We observe that on the news dataset, adding the Full UNT Lexicon slightly improves the F_1 score, while surface features alone don't give any improvement. 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 0.78. 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.

7 Model analysis

In order to better understand the obtained models, we visualized the decision trees as hierarchical diagrams, produced in the output of CLUS [15] . To ensure better interpretability of the models, we have constructed them in the following way: using a 10% pruning and 10%

testing dataset, we have used the F-test stopping criterion for splitting nodes. A node was split only when the test indicated a significant reduction of variance inside the subsets at the significance level of 0.10. The tree was then pruned with reduced error pruning using the validation dataset.

For clarity, we have only attempted to interpret the models using the lexicon and surface features. Bag-of-words features were omitted, since they resulted in deep one-branch nodes, which are difficult to visualize.

```
full_unt_pos > 0.0
+--yes: [OBJ]
+--no:  renderlex_noun_sum_neg > 0.0
  +--yes: [NEG]
  +--no:  numcaps > 0.0386
    +--yes: renderlex_adjective_abs > 0.4069
    | +--yes: h1w5 > 0.0312
    | | +--yes: [POS]
    | | +--no:  [OBJ]
    | +--no:  renderlex_all_sum > 3.866
    |   +--yes: [OBJ]
    |   +--no:  h1w5 > 0.0833
    |     +--yes: [OBJ]
    |     +--no:  full_unt_neg > 0.0
    |       +--yes: [OBJ]
    |       +--no:  repeat_vowel > 0.0244
    |    +--yes: [POS]
    |    +--no:  numvowel > 0.3429
    |      +--yes: [OBJ]
    |      +--no:  renderlex_all_abs > 2.1249
    |        +--yes: renderlex_all_sum > 2.7152
    |        | +--yes: [OBJ]
    |        | +--no:  [NEG]
    |        +--no:  [OBJ]
    +--no:  [OBJ]
```

Figure 10. Model constructed from training on Spanish news data (JRC-ES).

Figure 10 shows the tree, constructed by training the lexicon and surface feature representation of the news dataset. It shows that lexicon indicators are closest to the root, covering the most examples. The negative sum of noun scores has proven to be a good indicator for negative sentiment, suggesting that nouns are the more sentiment-bearing words in the news domain. Also, capitalization plays an important role in the model. While it is most likely a proxy for appearance of named entities, it shows that subjective statements tend to have more capitalized phrases. Also, the presence of questions (denoted as *h1w5*) tended to indicate a positive sentiment.

```
numvowel > 0.3246
+--yes: numcaps > 0.8462
| +--yes: [POS]
| +--no:  renderlex_all_sum_neg > 0.2682
|   +--yes: [POS]
|   +--no:  numvowel > 0.3566
|     +--yes: [NEG]
|     +--no:  renderlex_adverb_sum_neg > 0.4899
|       +--yes: [POS]
|       +--no:  repeat_letter > 0.0588
|    +--yes: [POS]
|    +--no:  [NEG]
+--no:  renderlex_adverb_abs > 0.52
  +--yes: renderlex_adverb_abs > 0.5964
  | +--yes: [POS]
  | +--no:  [NEG]
```

```
      +--no:  negation > 0.0
        +--yes: repeat_letter > 0.0357
        | +--yes: [NEG]
        | +--no:  [POS]
        +--no:  full_unt_neg > 0.0
          +--yes: [NEG]
          +--no:  length > 27.0
      +--yes: renderlex_noun_abs > 4.4911
      | +--yes: sad_face > 0.0
      | | +--yes: [POS]
      | | +--no:  [NEG]
      | +--no:  [OBJ]
      +--no:  [POS]
```

Figure 11. Model, constructed from training on Spanish social media (Render_es).

Figure 11 shows the model, trained with a Spanish social media dataset. Here, the primary features were the number of vowels, capitalized characters, along with letter repetition, reflecting how sentiment is typically expressed in social media and other forms of informal communication. Also, adverbs were shown to be the most important sentiment-bearing words, along with presence of negation words and emoticons.

```
renderlex_adjective_sum > 0.1096
+--yes: senticnet > 15.509
| +--yes: renderlex_adverb_abs > 8.1989
| | +--yes: swn_posneg_ratio > 5.2202
| | | +--yes: [POS]
| | | +--no:  numpunc > 0.0313
| | |   +--yes: renderlex_pos_links > 8025.0
| | |   | +--yes: renderlex_adjective_sum > 1.1693
| | |   | | +--yes: [POS]
| | |   | | +--no:  [NEG]
| | |   | +--no:  [NEG]
| | |   +--no:  [POS]
| | +--no:  [POS]
| +--no:  numvowel > 0.2808
|   +--yes: renderlex_adjective_abs > 0.3998
|   | +--yes: [NEG]
|   | +--no:  [POS]
|   +--no:  swn_total_pos > 17.0
|     +--yes: [NEG]
|     +--no:  renderlex_noun_sum > 7.8051
|       +--yes: [POS]
|       +--no:  [NEG]
+--no:  senticnet > 27.085
  +--yes: [POS] [98.0]: 182
  +--no:  repeat_letter > 0.1193
    +--yes: senticnet > 13.511
    | +--yes: [POS]
    | +--no:  [NEG]
    +--no:  numpunc > 0.0306
      +--yes: repeat_letter > 0.0626
| +--yes: renderlex_neg_links > 317.0
      | | +--yes: swn_total_obj > 272.5
      | | | +--yes: repeat_letter > 0.1001
      | | | | +--yes: [NEG]
      | | | | +--no:  renderlex_adjective_abs >
[.. omitted for brevity ..]
      | | | +--no:  [POS]
      | | +--no:  [NEG]
      | +--no:  swn_total_neg > 16.75
      |   +--yes: [NEG]
      |   +--no:  [POS]
      +--no:  [NEG]
```

Figure 12. Model, constructed from training on English review data (PangLee).

Figure 12 shows the same model, trained on the movie review dataset. Here, almost the entire model is dominated by various lexicon features – total scores, absolute scores, positive-negative ratios. To a minor extent, surface features such as vowel and letter repetition appear.

```
numcaps > 0.0345
+--yes: senticnet_neg > 1.113
| +--yes: [NEG]
| +--no:  renderlex_adjective_sum_neg > 0.2178
|   +--yes: [POS]
|   +--no:  senticnet_neg > 0.084
|     +--yes: swn_total_neg > 3.0
|     | +--yes: [POS]
|     | +--no:  numcaps > 0.037
|     |   +--yes: [OBJ]
|     |   +--no:  [NEG]
|     +--no:  renderlex_all_abs > 1.5025
|       +--yes: senticnet_abs > 0.816
|       | +--yes: renderlex_adverb_sum > 0.8143
|       | | +--yes: [POS]
|       | | +--no:  swn_total_neg > 4.0
|       | |   +--yes: renderlex_adjective_sum > 0.0
|       | |   | +--yes: [NEG]
|       | |   | +--no:  [OBJ]
|       | |   +--no:  [OBJ]
|       | +--no:  [NEG]
|       +--no:  [OBJ]
+--no:  [OBJ]
```

Figure 13. Model, constructed from training on English news (JRC-en).

Figure 13 shows a similar picture than its Spanish counterpart in Figure 8, showing the importance of lexicon features, followed by surface features. In English, although all words were sentiment-bearing, adjectives and adverbs seem to be more informative, compared to nouns in Spanish.

Figure 14 shows the social media sentiment model for English. Here, lexicons seem to be the most indicative, followed by vowel repetition and proportion, presence of negation and capitalization. These models also demonstrate that in English, lexicon features tend to be closer to the root than in its Spanish counterparts. This could be explained either by the quality and coverage of lexicons for the respective language or even cultural differences, where the sentiment expression is present not only in the choice of words, but also in the capitalization, use of punctuation and phrasing.

```
senticnet_neg > 0.007
+--yes: numvowel > 0.2963
| +--yes: negation > 0.0
| | +--yes: [POS]
| | +--no:  renderlex_all_abs > 0.1811
| |   +--yes: [NEG]
| |   +--no:  [POS]
| +--no:  [NEG]
+--no:  swn_total_neg > 1.5
  +--yes: numcaps > 0.0439
  | +--yes: [POS]
  | +--no:  [NEG]
  +--no:  repeat_letter > 0.125
    +--yes: numpunc > 0.0299
    | +--yes: [POS]
    | +--no:  numcaps > 0.0368
    |   +--yes: [POS]
    |   +--no:  [NEG]
    +--no:  renderlex_all_sum > 0.1013
      +--yes: numvowel > 0.2727
      | +--yes: renderlex_all_sum > 0.419
      | | +--yes: renderlex_pos_links > 442.0
      | | | +--yes: numpunc > 0.044
      | | | | +--yes: [POS]
      | | | | +--no:  [NEG]
      | | | +--no:  renderlex_adjective_sum > 0.0949
      | | |   +--yes: [POS]
      | | |   +--no:  [NEG]
      | | +--no:  [POS]
      | +--no:  [POS]
      +--no:  [NEG]
```

Figure 14**.** A model, constructed from training on English social media (Render_en).

8 Conclusions

The obtained results confirm that social media content is the domain which benefits from external knowledge. Topic-specific lexicons can bring some minor improvement over general purpose lexicons, but the best-performing approaches use a combination of bag-of-words and lexicons training data. We reported improvement on two English datasets, especially on social media, which benefited significantly from pre-processing, surface features, as well as lexicons.

Moreover, having a two-layer model brings the most consistent performance across all domains and languages. In terms of comparison against state-of-the art studies, the best result on the Pang and Lee datasets scores at 0.90 F1, while ours was slightly lower at 0.88. However, on the news domain, our best approach even improves the performance on the JRC-EN dataset from the original authors' 0.65 to our result of 0.68 F1. On the other hand, the voting and meta-models did not show any improvement over the W+F model, and only improved the concatenation on some datasets, while performance was even reduced on the other datasets.

The analysis of the models shows that there are major differences between domains on which features are considered important: while news and review domains benefited from lexicons, surface features were important only in social media. On the other hand, both languages exhibited similar behavior across the same domains in news. By interpreting the models trained on social media we show that, for Spanish, surface features were more important than lexicons, while the opposite was observed for English.

This paper also demonstrates the feasibility of using machine translation to obtain a training corpus in another language, showing that the performance obtained for JRC-ES was the same as in the original version - JRC-EN. Other research [10] shows 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.

Acknowledgements

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.

[14] Kaster, A. and Siersdorfer, S. and Weikum, G.: Combining text and linguistic document representations for authorship attribution, SIGIR workshop: Stylistic Analysis of Text For Information Access, 2005

[15] D. Kocev, C. Vens, J. Struyf and S. Džeroski, *Ensembles of multi-objective decision trees*. In J. Kok, J. Koronacki, R. de Mántaras, S. Matwin, D. Mladenic and A. Skowron, editors, Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Proceedings. Lecture Notes in Computer Science, volume 4701, pages 624-631, Springer, 2007

[16] Bryll, R. and Gutierrez-Osuna, R. and Quek, F.: Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern recognition, vol 36., no.6., pp. 1291-1302, Elsevier, 2003

[17] Porter, M. F.: Snowball: A language for stemming algorithms, 2001

[18] Yang, Yiming and Liu, Xin: A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42-49, ACM, 1999