RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D 2.1.2

Final version of the opinion mining toolkit

Editor:	Delia Rusu, JSI
Author(s):	Delia Rusu, Tadej Stajner, Inna Novalija, Blaz Fortuna, JSI
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	31 October 2012
Actual Delivery Date:	31 October 2012
Suggested Readers:	developers working on WP4 – Diversity Toolkit, developers creating case study prototypes (WP5)
Version:	0.1
Keywords:	opinion mining, sentiment analysis, bias detection

Disclaimer

This document contains material, which is the copyright of certain RENDER consortium parties, and may not be reproduced or copied without permission.

In case of Public (PU):
All RENDER consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):
All RENDER consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):
The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement. However, all RENDER consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):
The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the RENDER consortium as a whole, nor a certain party of the RENDER consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	RENDER – Reflecting Knowledge Diversity
Short Project Title:	RENDER
Number and Title of Work package:	WP2 Diversity Mining
Document Title:	D 2.1.2 - Final version of the opinion mining toolkit
Editor (Name, Affiliation)	Delia Rusu, JSI
Work package Leader (Name, affiliation)	Delia Rusu, JSI
Estimation of PM spent on the deliverable:	11

Executive Summary

This deliverable presents the final version of the Opinion mining toolkit, and is comprised of two main parts. In the first part we present and evaluate an updated version of the sentiment analysis algorithm.

We address 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. Our findings 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.

Our findings have been published in the proceedings of the 15th International Multiconference "Information Society - IS 2012", Ljubljana, Slovenia and annexed to this deliverable [Annex A.1].

In the second part of the deliverable we look at the macro level opinions by analysing reporting styles of various news sources. The differences in reporting are assessed using automatic methods, by comparing produced articles. We focus on comparisons along the following dimensions: topics, events and vocabulary.

List of authors

Organisation	Author
JSI	Delia Rusu
JSI	Tadej Stajner
JSI	Inna Novalija
JSI	Blaz Fortuna

Table of Contents

List of Figures

Abbreviations

BoW	Bag-of-words model
SVM	Support Vector Machines
MNB	Multinomial Naïve Bayes
WF-SVM	SVM using the 2-layer model
WF-SVMSc	WF with scaling and centering
MDS	Multidimensional scaling

1 Introduction

In this deliverable we present the final version of the Opinion Mining Toolkit, comprising sentiment analysis and media bias algorithms.

We start by describing an updated version of the sentiment analysis algorithm presented in D2.1.1 [11], which we evaluate in different settings, focusing on three types of datasets: news articles, reviews and social media. Detecting sentiment in social media is particularly challenging. 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. We therefore use three main feature types for learning, investigate the influence of sentiment dictionaries in such a setting and experiment with two different learning models.

Sentiment analysis is a natural language processing task which aims to predict the polarity (positive, negative or neutral) of sentiment data published by users, 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 [12]. One way to ameliorate this problem is to construct a lexicon of sentiment-bearing words 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. Research shows [22] 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. The first part of this deliverable 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.

In the second part of the deliverable we look at the macro level opinions by analysing reporting styles of various news sources. The differences in reporting are assessed using automatic methods, by comparing produced articles. We focus on comparisons along the following dimensions: topics, events and vocabulary.

This deliverable is structured as follows: we start by describing sentiment dictionaries and present our approach to building domain-specific sentiment dictionaries in Section 2. In Section 3, we detail the improvements made to the sentiment analysis algorithm. As most of the work has been published in proceedings of the 15th International Multiconference "Information Society - IS 2012", we briefly highlight the main findings of the paper (which we annexed to this deliverable – see Annex A.1), and provide details on the sentiment learning model analysis. Section 4 of the deliverable is dedicated to the analysis of news source bias, while in Section 5 we present concluding remarks.

2 Sentiment Dictionaries

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 exist a number of sentiment lexicons in English [1,2], the representation of sentiment resources in other lexicons is less notable.

2.1 Related work

SentiWordNet [1] is the most known sentiment dictionary. It is based on the WordNet [3] lexical database and represents each WordNet synset (synonym set) *s* 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.

As the polarity of words depends on the topic domain, several approaches to building context-aware sentiment lexicons have been proposed. Lu et al. [4] describe an optimization framework which allows combining different information sources for learning such a lexicon. Their approach is also sensitive to the aspect in context (e.g. for a laptop review, a “large” battery is negative whereas a “large” screen is positive, battery and screen being the aspects). Jijkoun et al [5] propose a different style of approach, by starting from an existing lexicon (clues) and focusing it. They perform a dependency parsing on a set of relevant documents, resulting in triplets (clue word, syntactic context, target of sentiment) that represent the domain specific lexicon. Kanayama and Nasukawa [6] apply the idea of context coherency (same polarity tend to appear successively) to the Japanese language. Starting from a list of polar atoms (minimum syntactic structure specifying polarity in a predicative expression), they determine a list of domain specific words using the overall density and precision of coherency in the corpus.

Recently, sentiment lexicons have been developed for other languages such as Spanish [7]. The approach relies on utilizing manual or automatically labelled data already available for the English language and the multilingual sense-level alignments available for the WordNet lexicon. The authors use the Opinion Finder [2] lexicon which provides manual annotations of subjectivity and polarity at word level, and transfer the manual annotations onto the English WordNet, relying on SentiWordNet constraints. In order to find the corresponding word sense, they matched Opinion Finder polarity strength to the SentiWordNet sense with the highest polarity score.

In our implementation, we have used the following lexicons:

- SentiWordNet (English) [1];
- SenticNet (English) [8];
- UNT Spanish – medium [7];
- UNT Spanish – full [7];
- RenderLex (English and Spanish), as described in this section;
- RenderLexLinks (English), as described in this section with added positive and negative link counts.

2.2 Approach

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.

Sentiment dictionaries developed in Render project are domain specific lexical resources, which contain words, part of speech tags and the relevant sentiment scores. We have set Telecommunications as the domain of primary interest, and the corpus, used for dictionaries development, was composed out of Twitter comments. We have started with a number of positive and negative seeds for different part-of-speech words (adjectives, nouns, verbs). Sentiment dictionaries are built in English and Spanish languages.

As discussed in [9], there are a number of approaches to develop the sentiment dictionary:

- manual approach;
- dictionary based approach;
- corpus based approach.

In our research on developing sentiment dictionaries we were following the work of Bizau et al. [9]. In this 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 provide language diversity and sentiments for different parts of speech:

1. Starting with positive and negative seed lists for adjectives, nous and verbs, we have expended the initial seed lists with using information about word synonyms and antonyms from WordNet.
2. From English and Spanish corpus of documents obtain all context same and context opposite pairs of words. In this step we have parsed each Twitter comment to extract all adjectives, verbs, nouns and conjunctions between them. In the same way, as Bizau et al. [9] and other researchers [10], we have been looking at words with the same context (represented by "and", "or" and "nor" connections) and words with the opposite context (represented by "but", "yet" connections). A particular importance was given to the negation present in the sentence. Examples of detecting context same and context opposite adjectives are provided at Figure 1.

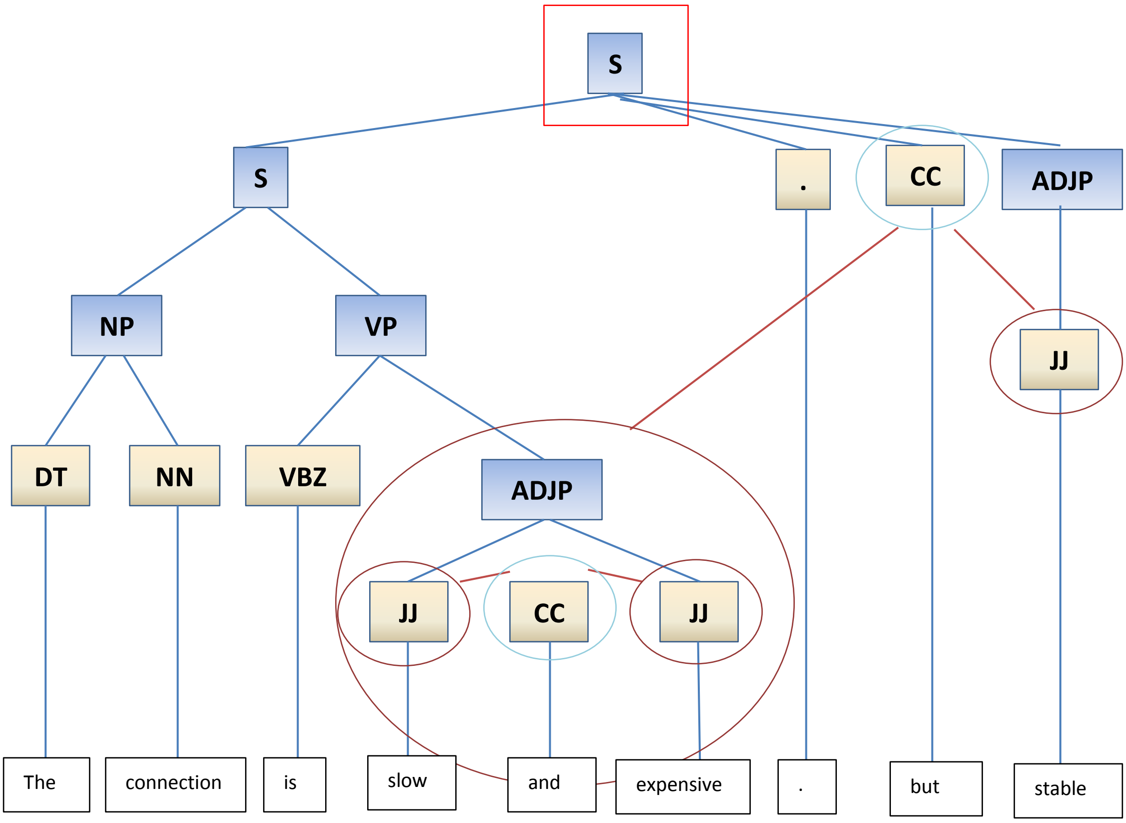


Figure 1. The parse tree and analysis of the sentence

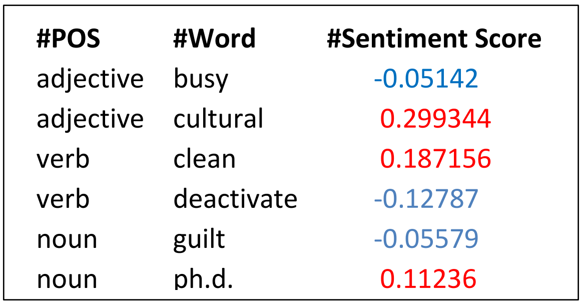
"The connection is slow and expensive, but stable".

3. The lists of context same and context opposite words have been pre-processed (the reflective relationships have been deleted). Furthermore, we have merged all noun tags into singular "NN" tag, all adjective tags into singular "JJ" tag, and finally, all verb tags into singular "VB" tag.
4. Using a list of context same and context opposite words, we have built a graph, starting with our word seed lists. Nodes in this graph represented words and edges represented the connections between the words.
5. In such way, we have created graphs for adjectives, nouns and verbs, in which the sign of the score for the particular word represented the word's orientation. A positive score characterizes a positive opinion orientation, while a negative score characterizes a negative opinion orientation.
6. In addition, besides part-of-speech tag and sentiment score, in our English dictionary we have provided several extra features, such as number of positive links and number of negative links for a particular word.

In contrast to Bizau et al. [9], we have created dictionaries not only in English, but also in Spanish. Our dictionaries were built not only for adjectives, but also for nous and verbs.

The produced English sentiment dictionary for the Telecommunication domain is composed out of around 2000 adjectives, 1700 verbs and 8000 nouns. The produced Spanish sentiment dictionary included around 650 adjectives, 2000 verbs and 4100 nouns.

Example words from the dictionary are presented below at Figure 2.



#POS	#Word	#Sentiment Score
adjective	busy	-0.05142
adjective	cultural	0.299344
verb	clean	0.187156
verb	deactivate	-0.12787
noun	guilt	-0.05579
noun	ph.d.	0.11236

Figure 2. Examples of words from sentiment dictionaries.

3 Sentiment Analysis using Sentiment Dictionaries

In this section 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 in improving our sentiment analysis algorithm (as described in D2.1.1 [11]).

Our findings have been published in the proceedings of the 15th International Multiconference "Information Society - IS 2012", Ljubljana, Slovenia [Annex A.1]. The paper describes two classification approaches to detecting sentiment in social media as well as news and reviews datasets. In this deliverable we summarize the main findings presented in the paper which is annexed to this deliverable. Moreover, we extend the paper by describing in more detail the model construction and analysis.

We start by briefly describing the feature construction, elaborate on the two models used for classification and present the model analysis.

3.1 Feature construction

We describe our data using three main feature sets: lexicon features, surface features and bag-of-words features. As we are dealing with multiple domains (social media, news and reviews), we have used different item types 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.

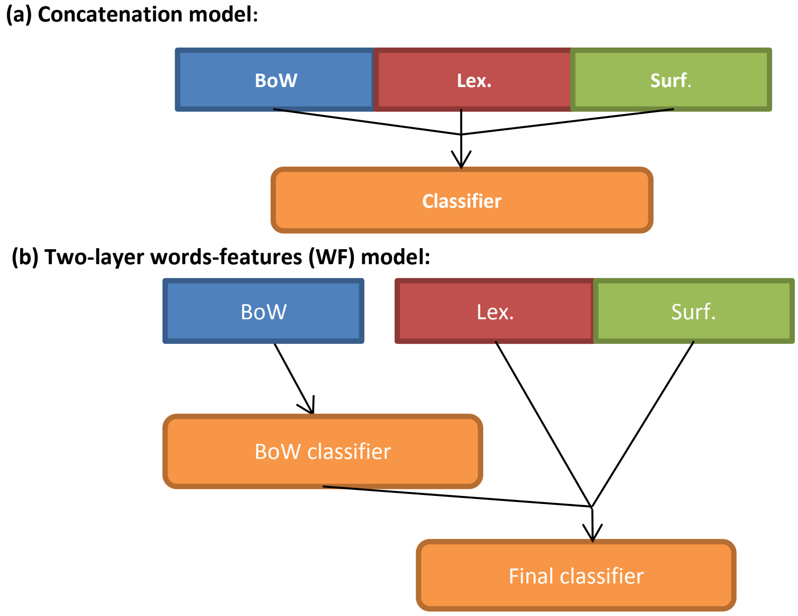
Bag-of-words features. We pre-process the textual contents by replacing URLs, numerical expressions and the names of opinions' targets with respective placeholders. We then tokenize this text, lower-casing and normalizing characters onto an ASCII representation, filtering for stop-words and weigh the terms using TF-IDF weights. The words were stemmed using the Snowball stemmer for English and Spanish. The punctuation is preserved.

Surface features. To accommodate social media, we have also used other text-derived features that can carry sentiment signal in informal settings, such as the count of fully capitalized words or the count of question-indicating words (for a full list of surface features, we refer the reader to our paper in [Annex A.1])

Lexicon features. We use lexicons in the form of features, where every word has assigned one or more scores, depending on the specificities of each lexicon [Annex A.1].

3.2 Models

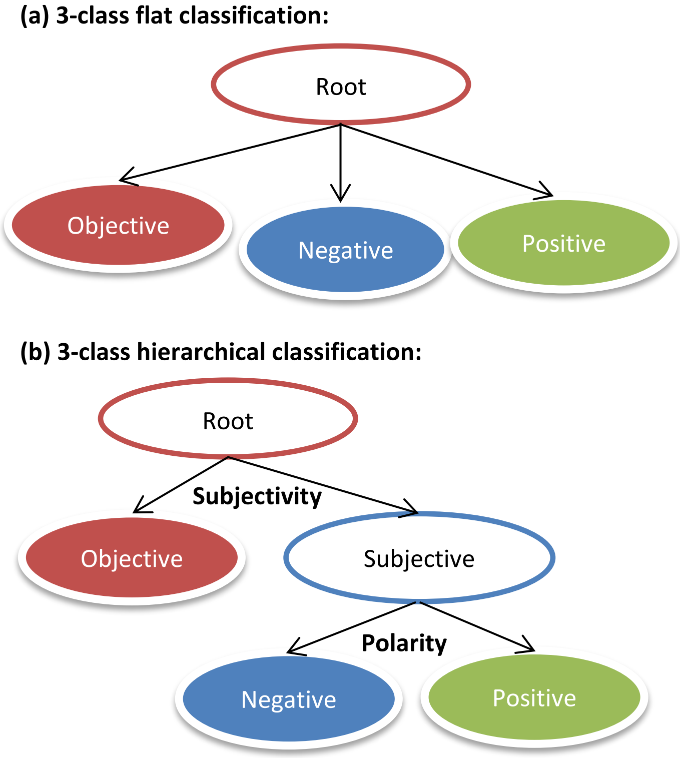
The data is composed of two modalities: bag-of-words features on one side, and having lexical and surface features, such as patterns and lexicon features on the other. In order to take differing distributions into account, we use two different approaches: either concatenating the features into a single features space, or using different models for each set of features. We compare the two modelling approaches, illustrated in Figure 3. 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.


(a) Concatenation model:

(b) Two-layer words-features (WF) model:

Figure 3. Diagrams of the (a) simple concatenation model and (b) the two-layer words-features model which encodes the BoW model output as features for the final model.

Sentiment analysis itself can be modelled either flat or hierarchically, as shown in Figure 4. In some domains, such as reviews, the problem can then be reduced to polarity classification, since all input data is inherently subjective. Furthermore, separating the sentiment problem into subjectivity and polarity has been shown to improve performance [12].


(a) 3-class flat classification:

(b) 3-class hierarchical classification:

Figure 4. Representations of sentiment classification as either (a) a standard three-class problem, or (b) a three-class hierarchical classification problem, composed of subjectivity classification and polarity classification.

3.3 Model analysis

In our evaluation setting we consider the following datasets:

- Pang & Lee review dataset, English [13];
- JRC news dataset, English [14];
- 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).

In order to better understand the obtained models (concatenation and the two-layer words-features), we visualize the decision trees as hierarchical diagrams, produced in the output of CLUS [22]. 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. As both the concatenation and the two-layer words-features models interpret the lexicon and surface features in the same way if we omit the BoW features (see Figure 3), we essentially analyse a single model.

```
full_unt_pos > 0.0
+--yes: [OBJ] [88.0]: 161
+--no:  renderlex_noun_sum_neg > 0.0
        +--yes: [SUBJ/NEG] [4.0]: 4
        +--no:  numcaps > 0.0386
                +--yes: renderlex_adjective_abs > 0.4069
                |       +--yes: h1w5 > 0.0312
                |       |       +--yes: [SUBJ/POS] [4.0]: 5
                |       |       +--no:  [OBJ] [5.0]: 6
                |       +--no:  renderlex_all_sum > 3.866
                |               +--yes: [OBJ] [21.0]: 32
                |               +--no:  h1w5 > 0.0833
                |                       +--yes: [OBJ] [10.0]: 17
                |                       +--no:  full_unt_neg > 0.0
                |                               +--yes: [OBJ] [4.0]: 8
                |                               +--no:  repeat_vowel > 0.0244
                |                                       +--yes: [SUBJ/POS] [2.0]: 4
                |                                       +--no:  numvowel > 0.3429
                |                                               +--yes: [OBJ] [113.0]: 129
                |                                               +--no:  renderlex_all_abs > 2.1249
                |                                                       +--yes: renderlex_all_sum > 2.7152
                |                                                       |       +--yes: [OBJ] [14.0]: 16
                |                                                       |       +--no:  [SUBJ/NEG] [9.0]: 14
                |                                                       +--no:  [OBJ] [43.0]: 47
                +--no:  [OBJ] [399.0]: 601
```

Figure 5. Model constructed from training on Spanish news data (JRC-ES).

Figure 5 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 (see Figure 5, the lines with parameter *h1w5*) tended to result in a positive sentiment.

```
numvowel > 0.3246
+--yes: numcaps > 0.8462
|       +--yes: [SUBJ/POS] [13.0]: 15
|       +--no:  renderlex_all_sum_neg > 0.2682
|               +--yes: [SUBJ/POS] [7.0]: 9
|               +--no:  numvowel > 0.3566
|                       +--yes: [SUBJ/NEG] [177.0]: 257
|                       +--no:  renderlex_adverb_sum_neg > 0.4899
|                               +--yes: [SUBJ/POS] [22.0]: 29
|                               +--no:  repeat_letter > 0.0588
|                                       +--yes: [SUBJ/POS] [20.0]: 32
|                                       +--no:  [SUBJ/NEG] [112.0]: 178
+--no:  renderlex_adverb_abs > 0.52
        +--yes: renderlex_adverb_abs > 0.5964
        |       +--yes: [SUBJ/POS] [10.0]: 19
        |       +--no:  [SUBJ/NEG] [8.0]: 8
        +--no:  negation > 0.0
                +--yes: repeat_letter > 0.0357
                |       +--yes: [SUBJ/NEG] [11.0]: 13
                |       +--no:  [SUBJ/POS] [12.0]: 17
                +--no:  full_unt_neg > 0.0
                        +--yes: [SUBJ/NEG] [8.0]: 10
                        +--no:  length > 27.0
                                +--yes: renderlex_noun_abs > 4.4911
                                |       +--yes: sad_face > 0.0
                                |       |       +--yes: [SUBJ/POS] [9.0]: 9
                                |       |       +--no:  [SUBJ/NEG] [2.0]: 2
                                |       +--no:  [OBJ] [15.0]: 22
                                +--no:  [SUBJ/POS] [75.0]: 102
```

Figure 6. Model, constructed from training on Spanish social media (RenderES).

Figure 6 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: [SUBJ/POS] [146.0]: 207
|       |       |       +--no:  numpunc > 0.0313
|       |       |               +--yes: renderlex_pos_links > 8025.0
|       |       |               |       +--yes: renderlex_adjective_sum > 1.1693
|       |       |               |       |       +--yes: [SUBJ/POS] [20.0]: 25
|       |       |               |       |       +--no:  [SUBJ/NEG] [28.0]: 53
|       |       |               |       +--no:  [SUBJ/NEG] [61.0]: 80
|       |       |               +--no:  [SUBJ/POS] [111.0]: 181
|       |       +--no:  [SUBJ/POS] [126.0]: 164
|       +--no:  numvowel > 0.2808
|               +--yes: renderlex_adjective_abs > 0.3998
|               |       +--yes: [SUBJ/NEG] [90.0]: 164
|               |       +--no:  [SUBJ/POS] [15.0]: 17
|               +--no:  swn_total_pos > 17.0
|                       +--yes: [SUBJ/NEG] [35.0]: 37
|                       +--no:  renderlex_noun_sum > 7.8051
|                               +--yes: [SUBJ/POS] [4.0]: 4
|                               +--no:  [SUBJ/NEG] [6.0]: 8
+--no:  senticnet > 27.085
        +--yes: [SUBJ/POS] [98.0]: 182
        +--no:  repeat_letter > 0.1193
                +--yes: senticnet > 13.511
                |       +--yes: [SUBJ/POS] [13.0]: 14
                |       +--no:  [SUBJ/NEG] [6.0]: 9
                +--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: [SUBJ/NEG] [40.0]: 45
                        |       |       |       |       +--no:  renderlex_adjective_abs > 3.0958
                        |       |       |       |               +--yes: [SUBJ/POS] [8.0]: 13
                        |       |       |       |               +--no:  renderlex_adverb_abs > 6.8693
                        |       |       |       |                       +--yes: renderlex_pos_links > 4737.0
                        |       |       |       |                       |       +--yes: renderlex_pos_links > 5239.0
                        |       |       |       |                       |       |       +--yes: [SUBJ/NEG] [78.0]: 99
                        |       |       |       |                       |       |       +--no:  renderlex_all_sum > 19.5557
                        |       |       |       |                       |       |               +--yes: [SUBJ/NEG] [6.0]: 7
                        |       |       |       |                       |       |               +--no:  [SUBJ/POS] [8.0]: 9
                        |       |       |       |                       |       +--no:  [SUBJ/NEG] [22.0]: 22
                        |       |       |       |                       +--no:  [SUBJ/NEG] [36.0]: 60
                        |       |       |       +--no:  [SUBJ/POS] [6.0]: 7
                        |       |       +--no:  [SUBJ/NEG] [30.0]: 32
                        |       +--no:  swn_total_neg > 16.75
                        |               +--yes: [SUBJ/NEG] [9.0]: 15
                        |               +--no:  [SUBJ/POS] [5.0]: 5
                        +--no:  [SUBJ/NEG] [94.0]: 161
```

Figure 7. Model, constructed from training on English review data (PangLee).

Figure 7 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: [SUBJ/NEG] [4.0]: 4
|       +--no:  renderlex_adjective_sum_neg > 0.2178
|               +--yes: [SUBJ/POS] [5.0]: 10
|               +--no:  senticnet_neg > 0.084
|                       +--yes: swn_total_neg > 3.0
|                       |       +--yes: [SUBJ/POS] [2.0]: 2
|                       |       +--no:  numcaps > 0.037
|                       |               +--yes: [OBJ] [120.0]: 135
|                       |               +--no:  [SUBJ/NEG] [3.0]: 7
|                       +--no:  renderlex_all_abs > 1.5025
|                               +--yes: senticnet_abs > 0.816
|                               |       +--yes: renderlex_adverb_sum > 0.8143
|                               |       |       +--yes: [SUBJ/POS] [1.0]: 2
|                               |       |       +--no:  swn_total_neg > 4.0
|                               |       |               +--yes: renderlex_adjective_sum > 0.0
|                               |       |               |       +--yes: [SUBJ/NEG] [3.0]: 4
|                               |       |               |       +--no:  [OBJ] [5.0]: 5
|                               |       |               +--no:  [OBJ] [70.0]: 74
|                               |       +--no:  [SUBJ/NEG] [3.0]: 3
|                               +--no:  [OBJ] [200.0]: 289
+--no:  [OBJ] [302.0]: 512
```

Figure 8. Model, constructed from training on English news (JRC-EN).

Figure 8 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.

```
senticnet_neg > 0.007
+--yes: numvowel > 0.2963
|       +--yes: negation > 0.0
|       |       +--yes: [SUBJ/POS] [2.0]: 2
|       |       +--no:  renderlex_all_abs > 0.1811
|       |               +--yes: [SUBJ/NEG] [5.0]: 5
|       |               +--no:  [SUBJ/POS] [1.0]: 2
|       +--no:  [SUBJ/NEG] [30.0]: 30
+--no:  swn_total_neg > 1.5
        +--yes: numcaps > 0.0439
        |       +--yes: [SUBJ/POS] [1.0]: 2
        |       +--no:  [SUBJ/NEG] [11.0]: 11
        +--no:  repeat_letter > 0.125
                +--yes: numpunc > 0.0299
                |       +--yes: [SUBJ/POS] [13.0]: 13
                |       +--no:  numcaps > 0.0368
                |               +--yes: [SUBJ/POS] [3.0]: 3
                |               +--no:  [SUBJ/NEG] [2.0]: 2
                +--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: [SUBJ/POS] [5.0]: 5
                        |       |       |       |       +--no:  [SUBJ/NEG] [2.0]: 2
                        |       |       |       +--no:  renderlex_adjective_sum > 0.0949
                        |       |       |               +--yes: [SUBJ/POS] [1.0]: 2
                        |       |       |               +--no:  [SUBJ/NEG] [10.0]: 10
                        |       |       +--no:  [SUBJ/POS] [6.0]: 7
                        |       +--no:  [SUBJ/POS] [7.0]: 7
                        +--no:  [SUBJ/NEG] [6.0]: 6
```

Figure 9. Model, constructed from training on English social media (RenderEN).

Figure 9 shows the social media sentiment model for English. Here, lexicons seem to be 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.

3.4 Discussion

The obtained results confirm that social media content is the domain which benefits the most from external knowledge. 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 reported improvement for 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 0.87. 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.

The analysis of the models, as presented in Section 3.3, 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 crucial in social media. On the other hand, both languages exhibited similar models 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.

We also demonstrate the feasibility of using machine translation to obtain a training corpus in another language, showing that the performance for JRC-ES was comparable to the original version - JRC-EN. Other research [10] shows promising approaches to facilitate the knowledge transfer via lexicons using specifically tailored machine learning approaches.

4 News Source Bias

In this section we look at the macro level opinions by analysing different reporting styles of various news sources. The differences in reporting are assessed using automatic methods, by comparing produced articles. We focus on comparisons along the following dimensions:

- Topic – what is being reported by each news source
- Events – what is the overlap of coverage between news sources
- Vocabularies – how similar (or different) is vocabulary used to describe same events.

To derive these comparisons we used 16 months of articles provided by Spinn3r [15], aligned the articles based on their content and applied several statistical learning techniques to derive and visualize similarities.

The reminder of this section is organized as follows. First, we present the data pre-processing and selection process for picking news sources. Then we describe how intersection is derived for each pair of news sources, followed by several ways to visualize these intersections [16].

4.1 Data pre-processing

The experiments are performed on a corpus containing 16 months of Spinner [15] outputs. The corpus covers the period between August 2008 and February 2010; all together 16 months of coverage. The corpus contains a combination of mainstream news articles, which are of interest in these experiments, and various social media sources such as blogs and micro blogs. The whole corpus takes around 5TB of space when uncompressed (1.7TB compressed).

In the first step, the corpus was filtered for mainstream news sources as identified by Spinner. The content of each article was cleaned [17], resulting in 80GB of clean text from around 40000 English feeds.

From various online directories we assembled a list of major news sources, covering most countries. The result is a database where each news source is annotated with its location (city and country). This list was matched against filtered Spinner data, to annotate aggregate feeds crawled via Spinner into sources and annotate them with location meta-data.

Finally, a list of sources was manually assembled for the experiment using the following criteria:

- **Broad coverage** – we wanted to cover most of the major countries of the world. This was difficult due to lack of available full English feeds for some parts of the world. Due to focus on English we also had a large bias in the number of available sources from USA and UK, which we tried to balance out in the selection processed.
- **Cover of mainstream news topics** – we wanted selected sources to emphasise coverage of daily, general news as well as world politics, economy and business news.
- **Sufficient availability of articles** – sources were required to have a significant enough presence in the corpus, in order to provide enough data for analysis.

This resulted in the list of sources presented in Figure 10. For each source we also list the amount of articles extracted from the corpus. It is important to note that blogs were not used in the experiments. This includes blogging sub-sites of mainstream media outlets, such as [http://www.nytimes.com/interactive/blogs/directory.html].

4.2 Topic analysis

Each source was classified into the DMOZ taxonomy using the Enrycher service [18]. This was done by individually classifying each articles, and then aggregating the counts of categories across all the articles. The results are presented in Figure 11, showing the top three categories for each source.

Source	Country	City	#articles
ABC News Australia	Australia	Sydney	32,929
Globe and Mail	Canada	Toronto	10,930
CBC Toronto	Canada	Toronto	18,896
Asia Times Online	China	Hong Kong	206
Hong Kong Standard	China	Hong Kong	582
People's Daily	China	Beijing	1,785
Fiji Times	Fiji	Suva	2,517
Euronews	France	Lyon	1,014
Deutsche Welle	Germany	Berlin	27,139
Maharashtra Times	India	Mumbai	13,093
Daily Express	Malaysia	Kota Kinabalu	1,458
Al Jazeera	Qatar	Doha	2,280
Southeast European Times	Serbia	Belgrade	1,203
Johannesburg Mail & Guardian	South Africa	Johannesburg	4,853
BBC News	United Kingdom	London	18,141
Daily Telegraph	United Kingdom	London	9,327
Times of London	United Kingdom	London	7,267
Chicago Sun-Times	USA	Chicago	6,231
Chicago Tribune	USA	Chicago	53,961
Boston Globe	USA	Boston	42,591
Los Angeles Times	USA	Los Angeles	35,070
New York Post	USA	New York City	3,059
New York Times	USA	New York City	230,850
San Francisco Chronicle	USA	San Francisco	19,809
USA Today	USA	Washington DC	48,130
Washington Post	USA	Washington DC	4,971
Las Vegas Sun	USA	Las Vegas	6,338
Zimbabwe Daily News	Zimbabwe	Harare	2,296

Figure 10. List of sources selected for analysis.

ABC News	Issues	Australia	New South Wales
Al-Jazeera	Warfare and Conflict	Issues	Specific Conflicts
BBC News	Issues	Warfare and Conflict	Specific Conflicts
Boston Globe	Publishing and Printing	Issues	Business / Investing
CBC	Society and Culture	North America / Canada	Issues
Chicago Tribune	Issues	United States / Illinois	Law / Law Enforcement
Fiji Times Online	Issues	Law Enforcement	News
India Times	Law / Law Enforcement	Issues	Society / Politics
Las Vegas Sun	Arts and Entertainment	Television	Law / Law Enforcement
Los Angeles Times	Issues	Los Angeles	United States / Presidents
New York Times	United States / Presidents	Issues	United States
Seattle Times	Issues	Government	Society / Government
Chicago Sun Times	Games	Video Games	Video Games / Action
The Globe and Mail	Computers / Internet	Business	Firms / Accountants
USA Today	Issues	United States	Society and Culture
Washington Post	Society and Culture	Society and Culture / Politics	Issues
Zimbabwe Telegraph	Arts / Literature	Arts / Literature	Communications
Daily Express	Issues	Asia / Malaysia	Society and Culture

Deutsche Welle	Law / Legal Information	Issues	Society / People
People.cn	Regional / Asia / China	Issues	Social Sciences / Economics
Euronews	Government / Multilateral	Society / Government	Issues
Mail & Guardian	Issues	Regional / Africa	Africa / South Africa
Telegraph	Issues	Arts	Society and Culture
The Standard	Business / Investing	Issues	Asia / China
The Times	Issues	Top	Financial Services

Figure 11. Top categories for news sources.

The results reflect in part the selection process indicated by a large “Issues” category presence. For the rest, we can see the regional bias with respect to covered topics (“Warfare and Conflict” for Al-Jazeera ad BBS News, “Australia” for ABC News, etc.). There are also some odd discrepancies, such as “Games” being very prominent in Chicago Sun Times, which is due to large presence of technology and entertainment in their feeds. An interesting point is “Government / Multilateral” topic which appears for Euronews, showing its bias towards European Union related stories.

4.3 Intersection

The next two steps in the experiment require pair-wise alignment of news articles between news sources. For example, given New York Times and BBC News, we would like to identify pairs of articles, one from each source, which cover similar events. Figure 12 shows one such example.

[ABC News] **About 500 angry government supporters massed outside the administrative court.** Thailand's Constitutional Court has dissolved the country's ruling parties and banned the Prime Minister, Somchai Wongsawat, from politics for five years. Protesters occupying Thailand's airports have broken into celebration.Thailand is again without a known leader tonight with a ruling by the country's Constitutional Court dissolving the key parties in the ruling coalition and banning Mr Somchai's from politics…

[Deutsche Welle] **Thailand's PAD threatens new demonstrations over leadership row**. In Thailand, the People's Alliance for Democracy has threatened to resume demonstrations if a candidate close to exiled former Prime Minister Thaksin Shinawatra is put forward to lead the country. Political parties are trying to come up with a candidate to replace Prime Minister Somchai Wongsawat, who was stripped of his post by the country's top court earlier this week. Thailand's constitutional court also dissolved the three…

Figure 12. Example of aligned article pair.

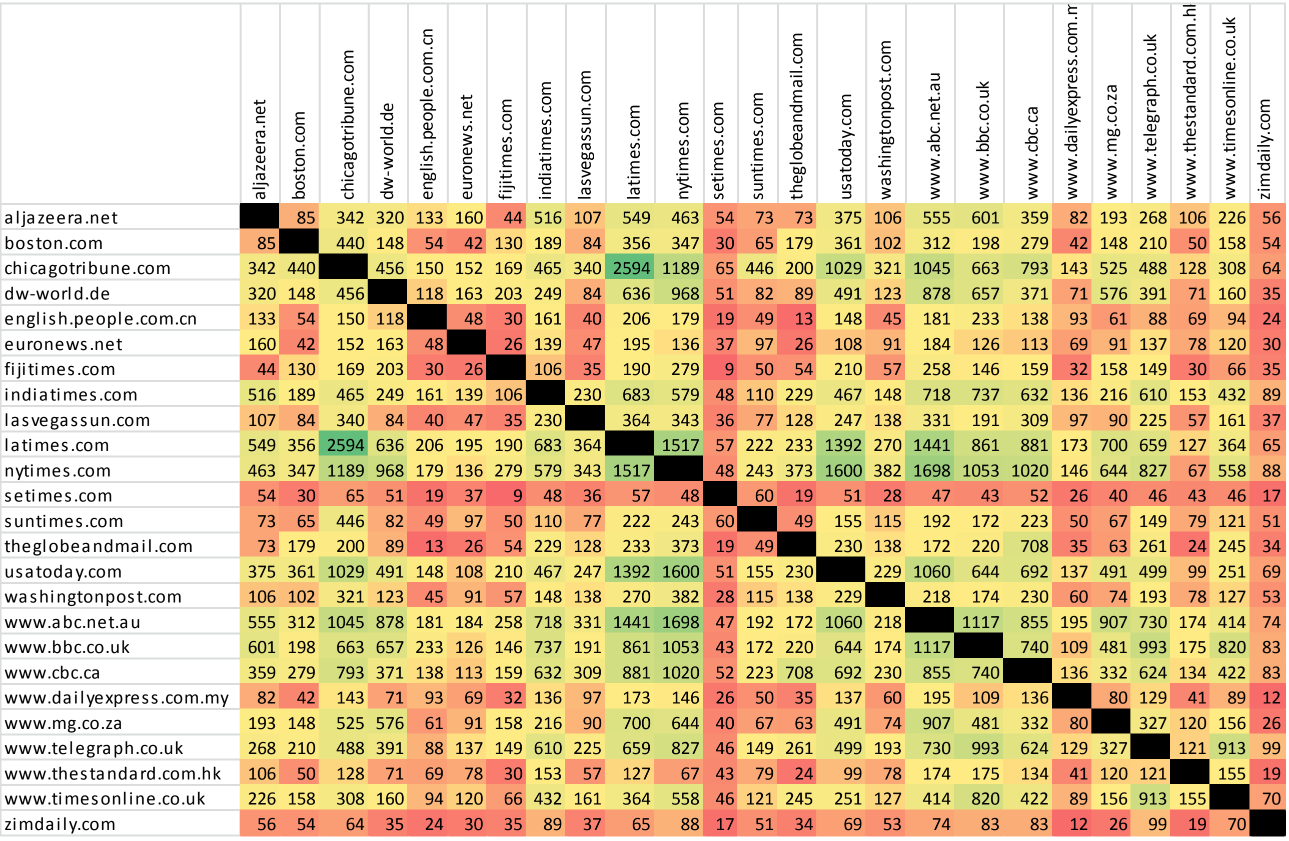
To align the articles we used the algorithm proposed in [17]. Let $S_1 = \{a_1, \dots, a_n\}$ and $S_2 = \{b_1, \dots, b_m\}$ be two news sources with their corresponding sets of articles. Each article is assigned a publish date $t(a)$. The alignment heuristic works as follows. For each article a we identify the most similar article $\hat{b}$ such that $\left|t(a) - t(\hat{b})\right| \leq T_p$. In the same way we identify the corresponding most similar article $\hat{a}$ for each article b. We use bag-of-words and cosine similarity to measure article similarity. Articles a and b are considered a pair if and only if $a = \hat{a}$ and $b = \hat{b}$, and $|t(a) - t(b)| \leq T_r$. In other words, articles are considered a pair if they are each other’s best match within the defined time window.

There are two time windows used in the heuristics. The first one, T_p, controls the size of the article pool from which the best match is selected. The bigger the window, the larger is the opportunity for an article to find a faraway match. In the experiments we set $T_p = 15$ days, based on previous experience and tests [17]. The second, T_r, defines the narrower time window around the event, making sure the articles are also chronologically similar. In the experiments we set $T_r = 2$ days, to compensate for any delays by crawler or publish cycles (e.g. morning in Australia vs. morning in USA).

We manually checked alignments of 100 articles and identified around 10 misalignments, resulting in precision of around 90%. It is important to stress that alignments degrade continuously. For example, one article might cover only a small sub-event covered in another. Also, the articles can overlap only partly, based on different focus of the reporting news source. We decided to still count such cases as alignment, since they can provide valuable insight into differences between sources.

The resulting intersection matrix is presented in Figure 13.

Based on the intersection counts we can compute the Jaccard similarity coefficient as the ratio between the intersection and union of articles from both sources. A coefficient of 1 would correspond to a complete overlap in coverage, while a coefficient of 0 would correspond to zero overlap. The result is presented in Figure 14, showing slight differences in distribution, compared to pure counts shown in Figure 13.



	aljazeera.net	boston.com	chicagotribune.com	dw-world.de	english.people.com.cn	euronews.net	fijitimes.com	indiatimes.com	lasvegassun.com	latimes.com	nytimes.com	setimes.com	suntimes.com	theglobeandmail.com	usatoday.com	washingtonpost.com	www.abc.net.au	www.bbc.co.uk	www.cbc.ca	www.dailyexpress.com.m	www.mg.co.za	www.telegraph.co.uk	www.thestandard.com.hl	www.timesonline.co.uk	zimdaily.com
aljazeera.net		85	342	320	133	160	44	516	107	549	463	54	73	73	375	106	555	601	359	82	193	268	106	226	56
boston.com	85		440	148	54	42	130	189	84	356	347	30	65	179	361	102	312	198	279	42	148	210	50	158	54
chicagotribune.com	342	440		456	150	152	169	465	340	2594	1189	65	446	200	1029	321	1045	663	793	143	525	488	128	308	64
dw-world.de	320	148	456		118	163	203	249	84	636	968	51	82	89	491	123	878	657	371	71	576	391	71	160	35
english.people.com.cn	133	54	150	118		48	30	161	40	206	179	19	49	13	148	45	181	233	138	93	61	88	69	94	24
euronews.net	160	42	152	163	48		26	139	47	195	136	37	97	26	108	91	184	126	113	69	91	137	78	120	30
fijitimes.com	44	130	169	203	30	26		106	35	190	279	9	50	54	210	57	258	146	159	32	158	149	30	66	35
indiatimes.com	516	189	465	249	161	139	106		230	683	579	48	110	229	467	148	718	737	632	136	216	610	153	432	89
lasvegassun.com	107	84	340	84	40	47	35	230		364	343	36	77	128	247	138	331	191	309	97	90	225	57	161	37
latimes.com	549	356	2594	636	206	195	190	683	364		1517	57	222	233	1392	270	1441	861	881	173	700	659	127	364	65
nytimes.com	463	347	1189	968	179	136	279	579	343	1517		48	243	373	1600	382	1698	1053	1020	146	644	827	67	558	88
setimes.com	54	30	65	51	19	37	9	48	36	57	48		60	19	51	28	47	43	52	26	40	46	43	46	17
suntimes.com	73	65	446	82	49	97	50	110	77	222	243	60		49	155	115	192	172	223	50	67	149	79	121	51
theglobeandmail.com	73	179	200	89	13	26	54	229	128	233	373	19	49		230	138	172	220	708	35	63	261	24	245	34
usatoday.com	375	361	1029	491	148	108	210	467	247	1392	1600	51	155	230		229	1060	644	692	137	491	499	99	251	69
washingtonpost.com	106	102	321	123	45	91	57	148	138	270	382	28	115	138	229		218	174	230	60	74	193	78	127	53
www.abc.net.au	555	312	1045	878	181	184	258	718	331	1441	1698	47	192	172	1060	218		1117	855	195	907	730	174	414	74
www.bbc.co.uk	601	198	663	657	233	126	146	737	191	861	1053	43	172	220	644	174	1117		740	109	481	993	175	820	83
www.cbc.ca	359	279	793	371	138	113	159	632	309	881	1020	52	223	708	692	230	855	740		136	332	624	134	422	83
www.dailyexpress.com.my	82	42	143	71	93	69	32	136	97	173	146	26	50	35	137	60	195	109	136		80	129	41	89	12
www.mg.co.za	193	148	525	576	61	91	158	216	90	700	644	40	67	63	491	74	907	481	332	80		327	120	156	26
www.telegraph.co.uk	268	210	488	391	88	137	149	610	225	659	827	46	149	261	499	193	730	993	624	129	327		121	913	99
www.thestandard.com.hk	106	50	128	71	69	78	30	153	57	127	67	43	79	24	99	78	174	175	134	41	120	121		155	19
www.timesonline.co.uk	226	158	308	160	94	120	66	432	161	364	558	46	121	245	251	127	414	820	422	89	156	913	155		70
zimdaily.com	56	54	64	35	24	30	35	89	37	65	88	17	51	34	69	53	74	83	83	12	26	99	19	70	

Figure 13. Number of articles in intersections.

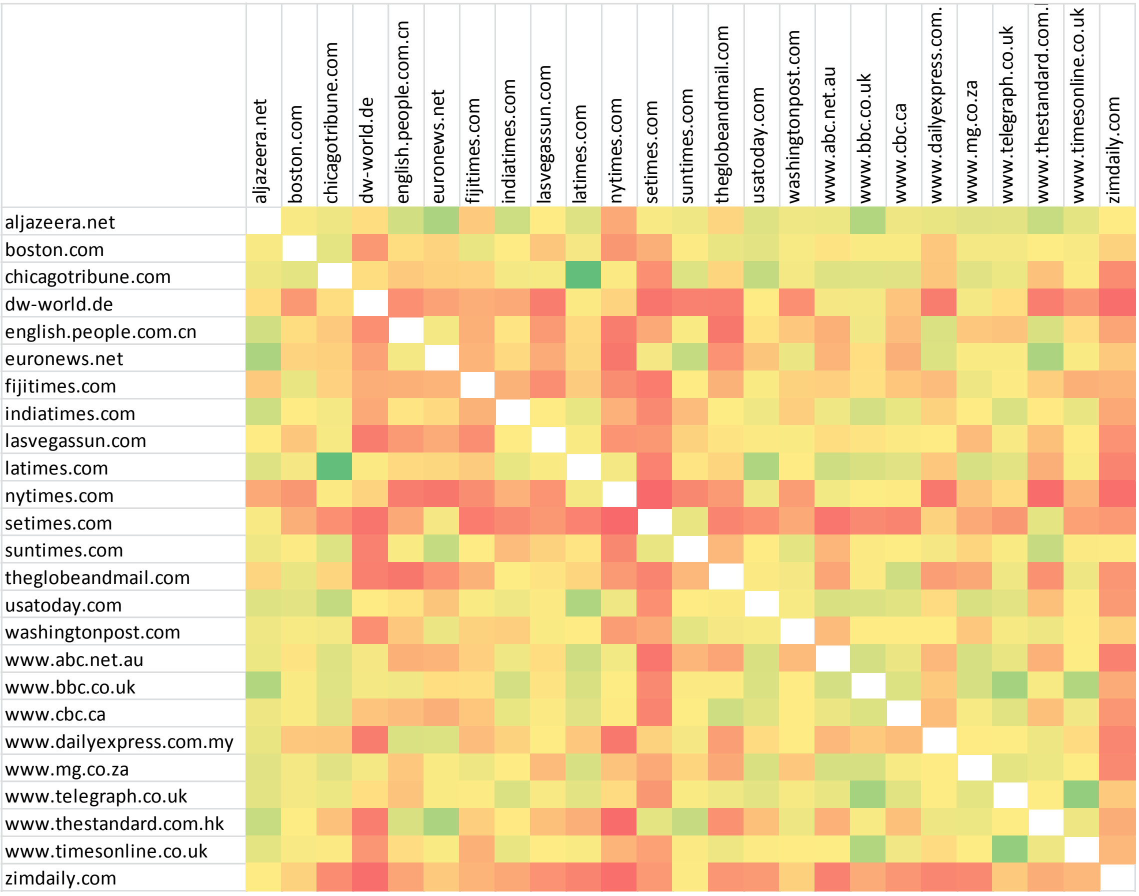


Figure 14. Colour coded Jaccard coefficient between sources (green is high).

Jaccard similarity coefficient matrix can be seen as a similarity matrix between sources. Using Multi-dimensional scaling (MDS) [19] we can embed the news sources onto a 2D plane such that more similar news sources are closer on the map, compared to less similar ones.

The result is shown in Figure 15. We can interpret the figure as a map of sources based on their coverage. There is a dense centre-right cluster, corresponding to large news sources, with either a large international coverage (e.g. bbc.co.uk) or larger usage of syndicated news (e.g. Associated Press). Moving to left-down we see Middle East (Al-Jazeera), European Union (Euronews) and China (Daily Express, People.cn). The map shows a bias towards Anglo-Saxon influence areas, which can be attributed to the focus on English writing news sources.

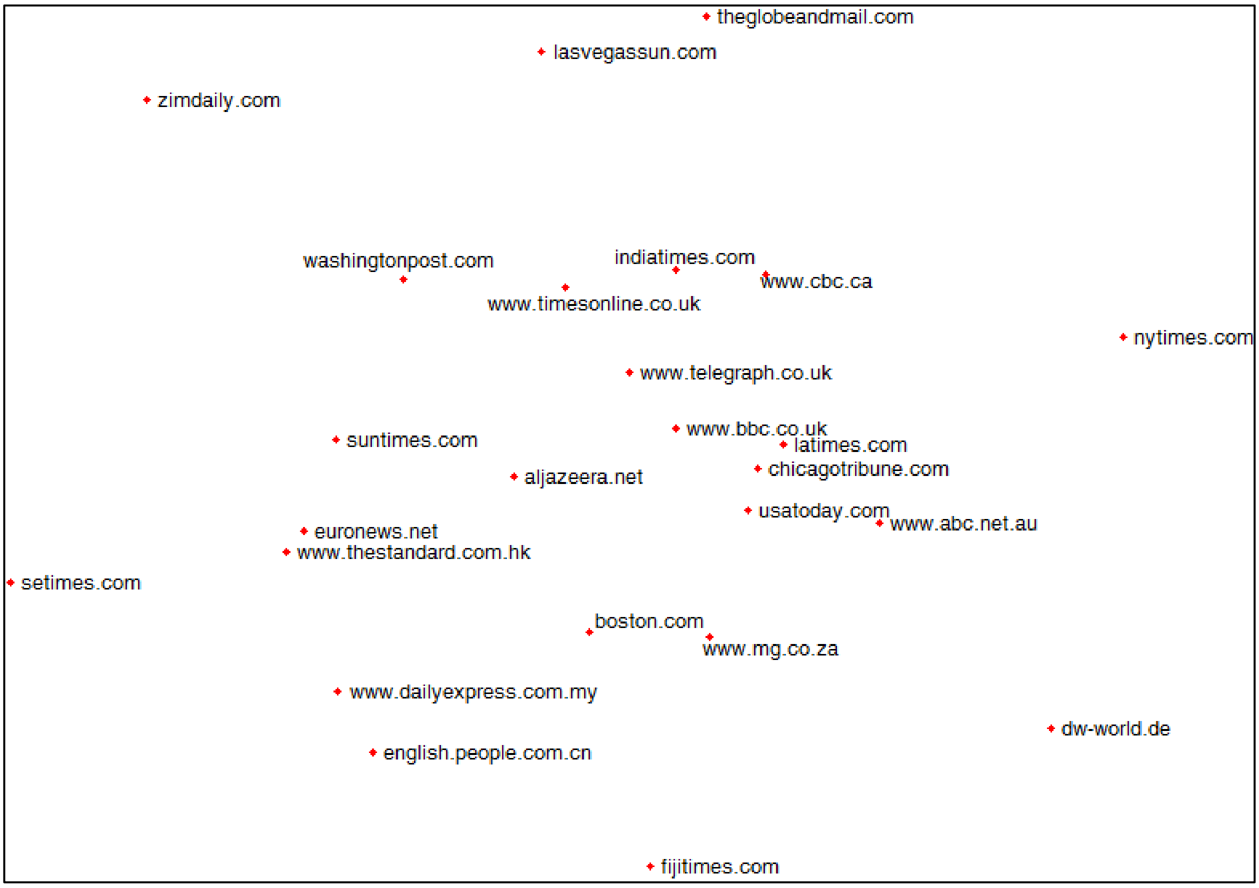


Figure 15. Map of sources positioned based on Jaccard similarity coefficient.

4.4 Classification accuracy

The final step is focused on a comparison of vocabulary similarity or difference when reporting about the same events. One way of estimating this is through the following simple experiment: we train a classifier, which can predict the source based on the article. For example, given the first article from Figure 12, can we correctly identify ABC news as the source? However, using all the articles from each source as part of the training data typically results in classifier identifying specific topics (such as ones in Figure 11), and not more subtle signals. We can avoid this by (a) creating a separate training corpus for each pair of news sources and (b) using only articles from the intersection. In this way we rule out the general topic bias, and the classifier must start relying on more subtle vocabulary level clues.

We used bag-of-words model to represent the article. This means that any results based on this can only be attributed to difference in vocabulary distributions between news sources. As part of this experiment we also identified words, which could provide unwanted signals to the classifier. For example, some news sources would have bias with respect to agencies they cite (e.g. AFP, AP, Reuters). Some other would start the article with the day of the week or month. We assembled such cases by observing popular words from the corpus, and created an extended list of stop-words used when defining the bag-of-words space.

Using the classification output, we can determine what are the most important keywords for distinguishing a particular news pair. Figure 16 shows an example of top keywords, based on their associated weight in the SVM weight vector, for a couple of pairs. As expected, Euronews tends to emphasize the European dimension, while New York Times emphasises the US dimension. Al-Jazeera would also tend to talk more about "violence", "Palestinians" and "Obama" compared to Euronews, which would emphasise more about "Israel" and "peace treaty talks".

Al-Jazeera	cent, president, told, Obama, protesters, Palestinian, government, minister, people,

	election, violence, polls, expected, officials, military, killed, economic, vote, fire
Euronews	today, countries, Iran, summit, supporting, parties, euro, European, percent, prime minister, state, militants, candidates, political, Israeli, full, Arabic, treaty, prime
Al-Jazeera	government, countries, told, president, cent, capital, people, attack, years, vote, securing, troops, ministry, parties, police, months, agency, heading, AIG, centre
New York Times	U.S, percent, United States, signed, china, bomb, UN, Taliban, billion

Figure 16. Top keywords distinguishing Al-Jazeera with Euronews (top) and New York Times (bottom).

The pair-wise training data, assembled for each news source pair, can also be used to estimate the classifier accuracy by doing a 5-fold cross validation with an SVM classifier [20]. We decided for accuracy due to a perfectly balanced training set. Accuracy of 100% indicates good separation of two news sources, indicating large difference in their vocabularies. Accuracy of around 50% indicates bad separation of two sources, indicating large similarity in their vocabularies. We used these two extremes to determine news source similarity, and devise a map using MDS. Given the news source similarities, MDS assigns each news source a point in the two-dimensional space. There result is presented in Figure 17. We can see the large UK influenced news sources in the centre. USA based news sources are largely grouped in several smaller clusters, placed around the map. Other sources typically occupy more isolated areas of maps (e.g. Al-Jazeera, Zimbabwe Daily).

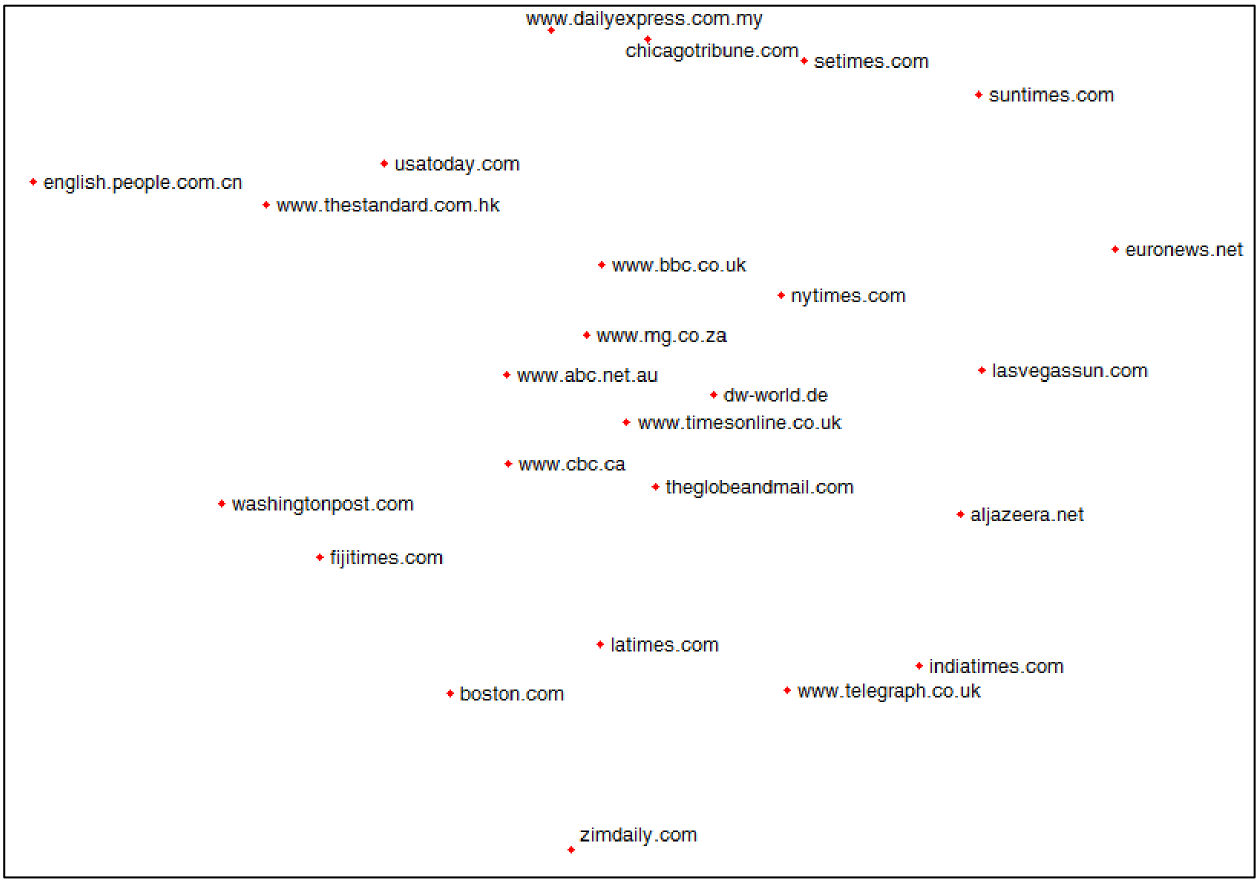


Figure 17. Map of sources positioned based on classification accuracy.

5 Conclusions and Future Work

In this deliverable we presented an updated version of the sentiment analysis algorithm and its evaluation on three different types of datasets: reviews, news and social media, in two languages: English and Spanish. Whereas the news and reviews data was already available for experimentation, we generated social media data for evaluation based on the datasets provided by our case study partner, Telefonica.

In our experiments we explored the influence of using background knowledge in the form of different sentiment lexicons, as well as the influence of various lexical surface features in improving our sentiment analysis algorithm. For this purpose, we created a domain-specific sentiment dictionary for the Telecommunications domain, in both English and Spanish.

Our results show that social media content is the domain which benefits the most from external knowledge, while topic-specific lexicons don’t bring improvement over general purpose lexicons. Two English datasets, especially social media, benefited significantly from pre-processing, surface features, as well as lexicons.

In future work we will explore cross-lingual learning, via approaches for training sentiment models using language resources from other languages.

The second part of this deliverable was dedicated to identifying bias in media, more precisely we looked at the macro level opinions by analysing reporting styles of various news sources. The differences in reporting are assessed using automatic methods, by comparing produced articles. We focused on comparisons along the following dimensions: topics, events and vocabulary.

References

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

[2] Wiebe, J. and Riloff, E. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Computational Linguistics, pages 486–497, Mexico City, Mexico.

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

[4] Lu, Y., Castellanos, M., Dayal, U., Zhai, C. Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach. In Proceedings of WWW 2011.

[5] Jijkoun, V., de Rijke, M. and Weerkamp, W. 2010. *Generating Focused Topic-Specific Sentiment Lexicons*. In Proceedings of the 48th Annual Meeting of the ACL.

[6] Kanayama, H. and Nasukawa, T. 2006. Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis. In Proceedings of the EMNLP.

[7] Perez Rosas, V., Banea, C., Mihalcea, R. 2012. Learning Sentiment Lexicons in Spanish. In Proceedings of the International Conference on Language Resources and Evaluations (LREC 2012), Istanbul, Turkey.

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

[9] 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.

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

[11] Rusu, D. (ed.). 2011. Prototype of the opinion mining toolkit. RENDER Deliverable D2.1.1.

[12] Pang, B., Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135.

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

[14] 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. 2010. Sentiment Analysis In the News. Proceedings of LREC.

[15] Spinn3r (retrieved 12.10.2012): http://spinn3r.com/company

[16] Fortuna, B., Galleguillos, C., Cristianini, C. 2008. Detecting the bias in media with statistical learning methods. Text Mining: Classification, Clustering, and Applications; edited by Ashok N. Srivastava, Mehran Sahami; Chapman & Hall/CRC press.

[17] Pasternack, J., and Roth, D. 2009. Extracting article text from the web with maximum subsequence segmentation. Proceedings of the 18th WWW conference, 2009.

[18] Rusu, D. (ed.). 2012. Prototype of the fact mining toolkit. RENDER Deliverable D2.2.1.

[19] Borg, I., Groenen, P.J.F., 2005. Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics). Springer.

[20] Cortes, C. & Vapnik, V. 1995. Support-vector network. Machine Learning, 20, 1–25.

[21] CLUS (retrieved 12.10.2012): http://dtai.cs.kuleuven.be/clus/index.html

[22] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics 37(2), pp 267-307.

Published papers

Informal sentiment analysis in multiple domains for English and Spanish

Stajner, T., Novalija, I., Mladenic, D. 2012. Informal sentiment analysis in multiple domains for English and Spanish. In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2012) co-located with the 15th International Multiconference on Information Society.

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

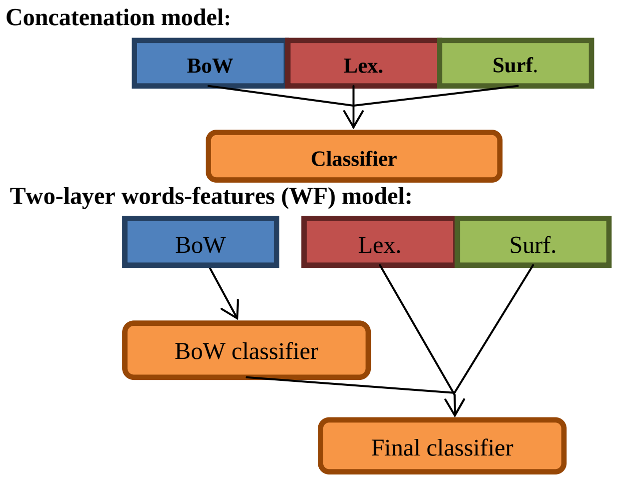


Figure 1: *Diagrams of the simple concatenation model and the two-layer words-features model which encodes the BoW model output as features for the final model.*

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.

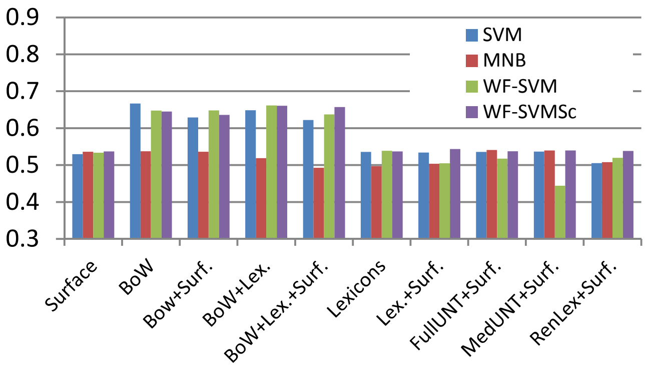


Table 1: Sentiment F_1 *scores on JRC-ES across settings.*

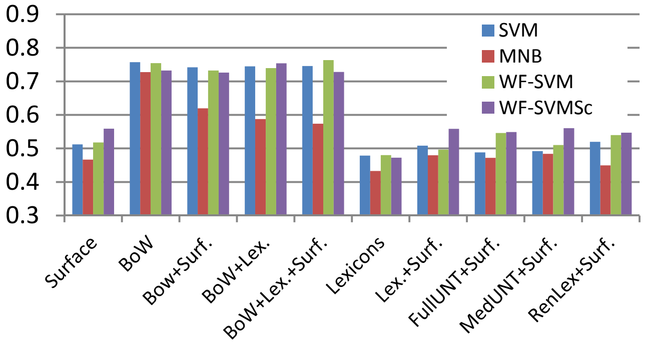


Table 2: Sentiment F_1 *scores on Render-ES across settings.*

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.

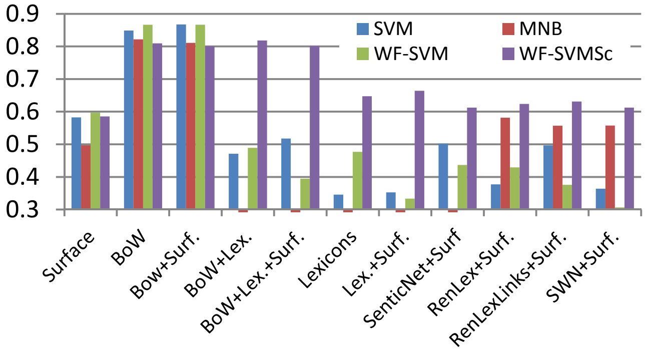


Table 3: Sentiment F_1 *score on PangLee across settings.*

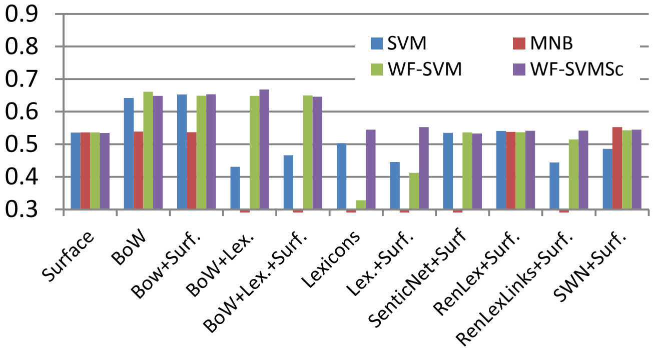


Table 4: Sentiment F_1 *scores on JRC-EN across settings.*

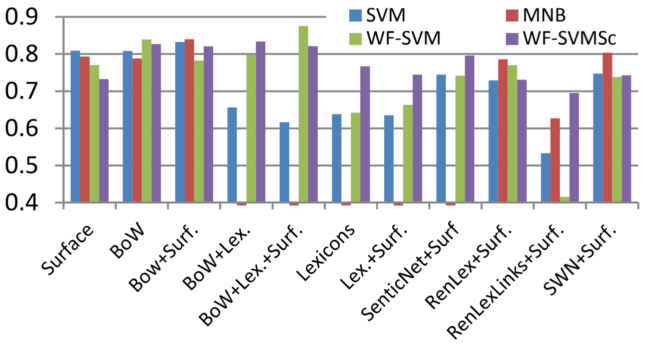


Table 5: Sentiment F_1 *scores on Render-EN across settings.*

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.