Lingmotif-lex: a Wide-coverage, State-of-the-art Lexicon for Sentiment Analysis

Antonio Moreno-Ortiz, Chantal Pérez-Hernández
University of Málaga
Spain
{amo,mph}@uma.es

Abstract

We present *Lingmotif-lex*, a new, wide-coverage, domain-neutral lexicon for sentiment analysis in English. We describe the creation process of this resource, its assumptions, format, and valence system. Unlike most sentiment lexicons currently available, *Lingmotif-lex* places strong emphasis on multi-word expressions, and has been manually curated to be as accurate, unambiguous, and comprehensive as possible. Also unlike existing available resources, *Lingmotif-lex* comprises a comprehensive set of contextual valence shifters (CVS) that account for valence modification by context. Formal evaluation is provided by testing it on two publicly available sentiment analysis datasets, and comparing it with other English sentiment lexicons available, which we adapted to make this comparison as fair as possible. We show how *Lingmotif-lex* achieves significantly better performance than these lexicons across both datasets.

Keywords: sentiment analysis, opinion mining, lexicons, language resources

1. Introduction

Sentiment Analysis has received increased attention in the last decade as a subtask of Natural Language Processing. The task can be roughly summarized as the classification of entire documents or parts of it (sentences or text segments that denote a certain aspect about a certain entity). Most efforts focus on the classification of domain-specific documents, usually short texts such as product reviews and tweets, although other more sophisticated tasks are becoming common, such as emotion and intensity detection. Traditionally, corpus-based and lexicon-based approaches have been distinguished in terms of the general system architecture. Strictly speaking, the former use a training corpus to extract textual cues found in each of the tagged documents, whereas the latter employ a sentiment lexicon where sentiment-carrying words are stored.

In practice, however, NLP practitioners combine methodologies from both approaches, for example, (Riloff et al., 2006). Generally speaking, lexicon-based approaches are preferred for sentence-level classification (Andreevskaia and Bergler, 2007), whereas corpus-based, statistical approaches are preferred for document-level classification.

The use of sentiment dictionaries is a widespread methodology, since it makes sense that the presence of certain sentiment-carrying words determine the polarity of the text in which they appear. WordNet (Fellbaum, 1998) has been a recurrent source of lexical information (Kim and Hovy, 2004; Hu and Liu, 2004; Andreevskaia and Bergler, 2006) either directly as a source of lexical information or for sentiment lexicon construction. Other common lexicons used in English sentiment analysis research include The General Inquirer (Stone and Hunt, 1963), MPQA (Wilson et al., 2005), and Bing Liu's Opinion Lexicon (Hu and Liu, 2004).

Yet other researchers have used a combination of existing lexicons or created their own (Hatzivassiloglou and McKeown, 1997; Turney, 2002). The use of lexicons has sometimes been straightforward, where the mere presence of a sentiment word determines a given polarity. However, negation and intensification can alter the valence or polarity of that word.[1] Modification of sentiment in context has also been widely recognized and dealt with by some researchers (Kennedy and Inkpen, 2006; Polanyi and Zaenen, 2006; Choi and Cardie, 2008; Taboada et al., 2011).

One disadvantage of relying solely on a sentiment lexicon is that different domains may greatly alter the valence of words, a fact well recognized in the literature (Aue and Gamon, 2005; Pang and Lee, 2008; Choi et al., 2009). A number of solutions have been proposed to these, mostly using ad hoc dictionaries, sometimes created automatically from a domain-specific corpus (Tai and Kao, 2013; Lu et al., 2011).

Lingmotif-lex has been embedded in the Lingmotif application (Moreno-Ortiz, 2017a), a fully user-focused, lexicon-based sentiment analysis system, since its availability in 2016. The Lingmotif application has been showcased at relevant conferences (Moreno-Ortiz, 2017b) and used as the main tool in a number of sentiment analysis shared tasks (Moreno-Ortiz, 2017c; Moreno-Ortiz and Pérez-Hernández, 2017). It currently supports English and Spanish; other languages (French and Italian) will be subsequently added and released.

The experience gained from participating in these sentiment and emotion classification tasks, as well as the input received from the application's users since its release, has served us to identify issues in the lexical resources, improve them, and refine them. The evaluation results shown in section 5. have provided us with the confidence to release a high-quality resource that can readily be used for real-world NLP tasks. Our intention, however, is to keep releasing improved versions of the resources.

Our Lingmotif sentiment analysis system is entirely lexicon-based. It implements a shifters system, which we describe in section 4. below, but it does not yet have entity or aspect management capabilities. This means that the quality of its results are entirely dependent on the quality

[1] The terms *valence* and *polarity* are used inconsistently in the literature. We use *polarity* to refer to the binary distinction positive/negative sentiment, and *valence* to a value of intensity on a scale.

of its lexical resources, that is, Lingmotif-lex, which is the object of our evaluation in this paper.

2. Creation Process

Lingmotif-lex is the result of many years of computational lexicography work and testing on many different sentiment analysis tasks. We started by merging available sentiment dictionaries. Specifically, we merged items from The Harvard General Inquirer (Stone and Hunt, 1963), MPQA (Wilson et al., 2005), and Bing Liu's Opinion Lexicon (Hu and Liu, 2004), which we reduced to a common format that included simply the lemma or form and its associated polarity. The resulting lexicon was then expanded semiautomatically by using a thesaurus and derivational generation rules. Since we decided to use a graded valence system instead of binary polarity, all items were manually ranked on a -5 to 5 scale by a team of trained annotators, using consensus and corpus linguistics techniques.[2]

The above-mentioned original resources, however, are characterized by their lack of attention to multiword expressions (MWEs). Multiword expressions not only denote polarity very often, but they also serve as good disambiguation resource in terms of polarity (Moreno-Ortiz et al., 2013), and are therefore key to successful lexicon-based sentiment analysis. In order to provide good coverage for MWEs, we used a number non SA-specific lexical resources, including common idioms from Wiktionary, which we tagged manually for valence. Ultimately, Lingmotif's lexicons are the result of intensive lexicographical work.

The obtained lexicon was further enhanced by correcting and adding lexical entries manually. A desktop application, Lingmotif[3] (Moreno-Ortiz, 2017a; Moreno-Ortiz, 2017b), was developed that enabled our team of annotators to easily analyze texts to identify errors and omissions. Since our aim was to create a domain-neutral sentiment lexicon, we chose a varied sample of texts for this development stage, with the aim of creating a resource that could be used on any sentiment analysis task, regardless of genre or topic.

To account for the domain specificity issue, the Lingmotif application allows the use of *plugin lexicons*. This is a flexible mechanism that allows users to develop and optionally apply a domain-specific lexicon for particular topics or domains. Lexical information contained in plugin lexicons overrides that in Lingmotif's core lexicon. When a plugin lexicon is selected for analysis, the plugin lexicon is searched first. If a word or phrase is found there, the core lexicon will not be searched for that item, and its information in the plugin lexicon will be used. Thus, plugin lexicons can be used to provide domain-specific sentiment items, but also to override polarity assignment in the core lexicon, for whatever reasons.

Plugin lexicons use exactly the same format as the core lexicon. In order to import a plugin lexicon, it must first be created as a UTF-8 encoded CSV file, which is then imported. Updating a plugin lexicon simply involves modifying the source CSV file and importing it again. Any number of plugin lexicons can be created in Lingmotif, but only one can be used for a given analysis.

Lingmotif-lex, along with a number of compatible plugin lexicons is ongoing work.

3. Format and Valence System

A *Lingmotif-lex* language set consists of three components: the lexicon itself, the context rules, which account for context modification of sentiment. A code library was developed to facilitate interfacing matching an input against the lexical resources. Table 1 summarizes the number of items contained in the current version of *Lingmotif-lex* for English.[4]

Item	**Count**
Single words (forms)	28,000
Multiwords (forms)	38,570
Emojis	130
Context rules (shifters)	700
Total	**67,400**

Table 1: Number of entries in *Lingmotif-lex*

The lexicon is stored in a plain text, tab-separated file encoded in UTF-8, which allows us to include Emojis just like any other lexical entry. Each lexicon entry consists of four data fields, which we describe in table 2.

Data field	**Example/List**
Word form	well-trained, looked_down_upon
Part of speech	[ALL, NN, JJ, VB, RB, UH, IN]
Polarity	[POS, NEG, NEU]
Intensity	[1, 2, 3]

Table 2: Lexicon format

In the past, polarity and intensity were expressed as a simple negative or positive integer. However, the current system allows us to express cases where polarity is not well defined or highly context-dependent, but where the presence of intensity is unquestionable, which is a very common situation. Words and expressions such as ”wild”, ”wicked”, ”sick”, or ”oh, my god” are clear examples. Thus, even if this format presents some more processing difficulty, it gains in expressive power, and allows users to apply their own disambiguation techniques if required.

The valence system in the release version, based on a 3-point intensity scale is also different from the scale used in previous versions used in the Lingmotif application, and described elsewhere (**?**), where a more fine-grained, 5-point scale was used. We found this to be useful for some cases, such as graded adjective series, but harder to define in many other scenarios. The current coarser 3-point scale is more intuitive for annotators (low, mid, high intensity), and just as useful for practical purposes.

All entries are lower case, and contain no blanks. The following are examples of *Lingmotif-lex* single-word entries:

```
kind JJ POS 2
```

[2] This scale was kept for some time, but the final version was reduced to a 3-point intensity scale. The valence system is described in section 3..

[3] The Lingmotif application is available for download at http://tecnolengua.uma.es/lingmotif

[4] These figures might be different in the final release version.

```
corrupt ALL NEG 2
curious ALL NEU 1
```

Single-word entries are pos-tagged as ALL by default, meaning that the word, functioning as any part of speech, has that valence. When a word takes the valence only when functioning as a particular part of speech, the part of speech is given (using the Penn Treebank tag set). Part-of-speech data is used by the context rule matching system.

Multiwords are always pos-tagged, and otherwise treated just like single words, the only difference being signaled by the presence of underscores separating the individual words that make them up. English phrasal verbs are also treated as MWEs. Multiword expressions may consist of up to six words, some of which may be placeholders, marked by ”0”. For example, the following set of entries:

```
brought_into_harmony VB POS 2
brought_0_into_harmony VB POS 2
brought_0_0_into_harmony VB POS 2
brought_0_0_0_into_harmony VB POS 2
```

would match the strings ”brought into harmony”, ”brought them into harmony”, ”brought the situation into harmony”, and ”brought the guests back into harmony”. This system allows to express and effectively match the vast majority of multiword expressions. Lingmotif generates such sets of entries from a more user-friendly format with the form

```
<bring>_3_into_harmony VB POS 2
```

where the word in angled brackets means it is a lemma, so all its forms hould be generated during import, and the integer in braces specifies the number of words that may occur between both part of the multiword expression, which generates the forms shown above.

Conceptually, however, there is an important difference between single and multi words in *Lingmotif-lex*. Most single words (except polarity-ambiguous ones) have a non-neutral polarity tag, and invariably have an intensity greater than 0, the assumption being that any single word not in the lexicon is, generally, not a sentiment word. On the other hand, the lexicon includes multiwords which may not be sentiment-carrying but contain individual words that are. The rationale behind this is that, were they not included in the lexicon, those individual words would be matched and (wrongly) identified as sentiment items (for example the word ”kill” in ”kill time”. Further explanations and examples are provided in Moreno-Ortiz et al. (2013).

4. Sentiment shifters

Real-world NLP tasks involving sentiment analysis have often relied on sentiment lexicons, but less so on a wide-coverage system that accounts for contextual modification of the lexical items. Context can greatly modify the sentiment of, fundamentally, any sentiment word, either by inversion, such as in negation, where the overall polarity is inverted; or by intensification and downtoning by means of, for example, a quantifier. This means that the sheer presence of a sentiment word does not determine the polarity or valence assigned to that sentiment word in the lexicon. A shifters system may help accounting for contextual modification of sentiment, and is particularly interesting for fine-grained sentiment analysis, such as aspect-based SA, where the text's overall classification/score is not enough (Yu et al., 2016). However, such systems have rarely been fully implemented.

Lingmotif-lex does include a with a wide-coverage shifters system, which we have implemented by means of context rules, basically a template matching system on the input text. We use a similar linguistic approach to Polanyi and Zaenen (2006), Kennedy and Inkpen (2006), or Taboada et al. (2011). Our context rules work by specifying words or phrases that can appear in the immediate vicinity of the identified sentiment word. Our set of context rules has been compiled through extensive corpus-based work and have been tested on several sentiment analysis datasets. Currently, *Lingmotif-lex* contains over 700 such rules.

Table 3 shows the data structure used by *Lingmotif-lex*'s shifters system.

Data field	**Example/Value list**
Part of speech	[NN, JJ, VB, RB]
Polarity	[+, -]
Shifter form	not, tremendous*, pretty_much
Location	[L, R, LR]
Span	[1, 2, 3, 4, 5]
Result	[INT*n*, DOW*n*, VAL*n*, INV0]

Table 3: Context rules format

As can be deduced from the example shifters in table 3, they can be either a literal string or contain the asterisk as a wildcard (e.g., ”tremendous*” matches both the adjective and the adverb ”tremendously”). Shifters can also be multiword expressions, in which case they must be included in the lexicon. The *location* data field expresses the position of the shifter with reference to the sentiment word, and *span* defines the maximum number of words from it. The *result* values (INT*n*, DOW*n*, VAL*n*, INV0) define both the type of shifting produced (intensification, downtoning, value, inversion) and the degree of the shift (as expressed by the integer *n*).

Table 4 provides examples of all types of sentiment shifters according to the effect they produce on the resulting text segment.

Shift type	**Example context rule**

Intensification	`JJ +- highly L 1 INT2`
	`VB +- literal* LR 2 INT2`

Downtoning	`JJ +- sort_of L 1 DOW1`
	`NN +- limited L 1 DOW1`

Inversion	`NN +- decreas* LR 3 INV0`
	`JJ +- never L 2 INV0`

Final value	`VB + try_to L 1 VAL0`
	`JJ +- would_have L 3 VAL0`

Table 4: Context rules examples

The set of rules is in a separate plain text, CSV file, and a Python library provides easy access to it, allowing users to programmatically choose to individually apply inversion, intensification, downtoning, or no rules at all, in which case the original valence of the lexical items will be returned. The library returns the list of matched *sentiment items* and their corresponding valences. Sentiment items, after applying context rules, may be a single word form, a sequence of word forms matching a multiword expression, or a sequence of word forms matching a single or multiword item and the strings that make up the context rule.

5. Evaluation methodology

We evaluate *Lingmotif-lex* in terms of performance on a typical sentiment classification task, using two publicly available sentiment analysis datasets, and compare it with five other well-known available English sentiment lexicons. Comparing sentiment lexicons is not easy, since different resources have different formats and data. For example, they may or may not include part of speech information, and, if this is present, different tagsets may be used. Their entries may be word forms or lemmas, and the valence system may also be different, from simple polarity to fine-grained data.

5.1. Datasets

After many years of research in sentiment analysis, the types of documents that have been dealt with in the field are varied. However, two types stand out among the rest: user reviews of products or services, and microblogging short texts, particularly Twitter data.

For our performance evaluation we chose two highly representative datasets of these two types. For the first, we used the classic movie reviews dataset from Pang and Lee (2004)'s seminal paper. This database has been widely used in the literature both for training and testing algorithms and resources. It is made up of 1,000 positive and 1,000 negative movie reviews tagged for polarity (pos/neg).[5] For Twitter data, we chose a dataset specifically designed for testing sentiment analysis techniques on microblogging texts, STS-Gold (Saif et al., 2013), which contains 2,034 tweets tagged for polarity (1,402 negative, 632 positive).

Present-day SA shared tasks commonly aim to classify texts into finer categories, including the neutral category and degrees of intensity, but we think simpler datasets, such as these two, provide a solid base on which to test sentiment lexicons, since there is less chance of annotation issues.

5.2. Sentiment lexicons

We compared the performance of *Lingmotif-lex* with five other well-known sentiment lexicons:

1. The General Inquirer (Stone and Hunt, 1963). This is one of the oldest sentiment lexicons publicly available. It is based on work in cognitive psychology and content analysis. This resource offers syntactic, semantic, and pragmatic information to part-of-speech tagged words, with 1915 positive and 2291 negative words. Lexical items for ”yes” and ”no” (in the sense of refusal) are grouped in separate categories and further semantic dimensions, such as strength or active/passive orientation, are also included[6].

2. Bing Liu Opinion Lexicon (Hu and Liu, 2004). A compilation of about 6,800 words drawn from product reviews, originally labeled using a bootstrapping method using WordNet adjective synsets and their antonyms (Hu and Liu, 2004). It contains 2006 positive and 4783 negative words[7].

3. MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon (Wilson et al., 2005). This resource contains 2,718 positive and 4,912 negative words drawn from a combination of sources, including the General Inquirer lists, the output of the system created by Hatzivassiloglou and McKeown (1997) and a bootstrapped list of subjective clues (Riloff and Wiebe, 2003), hand-labeled for sentiment. The lexicon also includes labels for reliability (strongly subjective or weakly subjective) and four polarity tags: positive, negative, both and neutral. The majority of words are marked as having either positive (33.1 percent) or negative (59.7 percent) polarity, whereas only a small number of clues (0.3 percent) are marked as having both positive and negative, and 6.9 percent of the clues in the lexicon are marked as neutral[8].

4. SentiWordNet 3.0 (Baccianella et al., 2010). a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications (Baccianella et al., 2010). It is the result of automatically annotating all WordNet synsets according to their degrees of positivity, negativity, and neutrality. The current version is 3.0 (based on WordNet 3.0) and differs from previous versions in the algorithm used for annotation, which now refines the scores by using a random-walk step in addition to the initial semi-supervised learning step[9].

5. Sentiment140 Lexicon (Mohammad et al., 2013). Created from a collection of 1.6 million tweets that contain positive and negative emoticons. This is the only lexicon in this list that contains multiwords, including 62,468 unigrams, 677,698 bigrams, and 480,010 pairs tagged as either positive or negative[10].

We made every effort to adapt all resources to a common usable format, although, inevitably, some information was lost in some adaptations. For example, some lexicons, such as MPQA, list lemmas rather than forms (implicitly, in this case), so we generated all forms for each lemma, which surely generates some forms that do not exhibit the same polarity.

5.3. Training and testing procedure

All lexicons were evaluated using the same metrics (precision, recall, and F1 score) obtained with the same training and testing procedure. Each dataset was split using a

[5] The dataset is available at http://www.cs.cornell.edu/people/pabo/movie-review-data/

[6] Available at http://www.wjh.harvard.edu/ inquirer/.

[7] Available at http://www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar

[8] Available at http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.

[9] Available at http://sentiwordnet.isti.cnr.it/.

[10] Available at http://saifmohammad.com/WebPages/lexicons.html.

widely-used approach to sentiment classification at the document level: 80 percent for training and 20 percent for testing. Sentiment words present in each of the lexicons were searched in the texts, and a positive and negative score was calculated for each text as the sum of the valences found in the lexicon. When the lexicon simply contained polarity, a valence of *1* was added to that polarity score; when the lexicon included a more fine-grained valence specification, that number was used. The classifier was trained exclusively on these two features: positive score and negative score, calculated as the accumulated sum of the valences of the matching items, as specified in each of the lexicons.

As for the classifier itself, we used a ”traditional” support vector machine algorithm, specifically the SVC implementation in the Python-based scikit-learn (Pedregosa et al., 2011) machine learning toolkit, with the default RBF kernel and parameters for all lexicons.

This methodology minimally guarantees that the best results would be obtained by the best lexicons, since all other conditions where kept equal in all cases, the lexicon being the only variable. It is of course debatable whether a larger and/or more varied sample should be used, but we think these two datasets are fairly representative of typical sentiment analysis tasks.

6. Evaluation results

Tables 5, 6, 7, 8, 9, and 10 show the performance results we obtained for each of the lexicons on both datasets. It is interesting to see how all lexicons exhibit significant differences for each of the classes (positive, negative). Negativity consistently obtains high precision but low recall, whereas the positive class is just the opposite (better recall, poorer precision). This is surely due to the fact that negativity is expressed using more sophisticated linguistic resources (irony, sarcasm, understatements, etc.), and, therefore, not so easily identified. The difference in performance among classes is particularly evident in some lexicons, such as SentiWordNet (8) and Sentiment 140 (9).

Movie reviews				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.81	0.64	0.71	1278
pos	0.54	0.74	0.62	722
avg/total	0.71	0.68	0.68	2000
STS-Gold				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.96	0.73	0.83	1842
pos	0.21	0.69	0.32	192
avg/total	0.89	0.73	0.78	2034

Table 5: Evaluation results - General Inquirer

Movie reviews				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.78	0.70	0.73	1115
pos	0.66	0.75	0.70	885
avg/total	0.73	0.72	0.72	2000
STS-Gold				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.87	0.84	0.85	144
pos	0.63	0.69	0.66	585
avg/total	0.80	0.80	0.80	2034

Table 6: Evaluation results - Bing Liu

Movie reviews				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.75	0.69	0.72	1092
pos	0.66	0.73	0.69	908
avg/total	0.71	0.71	0.71	2000
STS-Gold				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.90	0.77	0.83	1628
pos	0.41	0.64	0.50	406
avg/total	0.80	0.75	0.76	2034

Table 7: Evaluation results - MPQA

Movie reviews				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.97	0.71	0.82	1901
pos	0.13	0.63	0.22	133
avg/total	0.91	0.71	0.78	2034
STS-Gold				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.97	0.71	0.82	1901
pos	0.13	0.63	0.22	133
avg/total	0.91	0.71	0.78	2034

Table 8: Evaluation results - SentiWordNet

Movie reviews				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.73	0.64	0.68	1149
pos	0.58	0.69	0.63	851
avg/total	0.67	0.66	0.66	2000
STS-Gold				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.94	0.76	0.84	1733
pos	0.34	0.71	0.46	301
avg/total	0.85	0.75	0.78	2034

Table 9: Evaluation results - Sentiment 140

Tables 11 and 12 summarize the results we obtained for the movie reviews dataset and the STS-Gold dataset, respectively, for each of the lexicons.

As these results show, Lingmotif-lex obtains a higher score on both datasets, followed by Bing Liu’s lexicon, which it improves by 0.3 and 0.4 respectively, a significant difference given the tight scores. This consistency of results is desirable, since it is a sign of reliability and predictability, something that may not be said of the other lexicons, especially MPQA, which ranked third on the movie reviews dataset, but last on the STS-Gold dataset.

7. Conclusions

In this paper we have presented *Lingmotif-lex*, a wide-coverage, non-domain-specific lexicon for sentiment analysis in English. We have described the creation process of

Movie reviews				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.81	0.72	0.76	1124
pos	0.68	0.78	0.73	876
avg/total	0.75	0.74	0.75	2000
STS-Gold				
Class	**Precision**	**Recall**	**F1**	**Support**
neg	0.93	0.85	0.89	1547
pos	0.62	0.81	0.70	487
avg/total	0.86	0.84	0.84	2034

Table 10: Evaluation results - Lingmotif-lex

Lexicon	Precision	Recall	F1 Score
Lingmotif-lex	0.75	0.74	0.75
Bing Liu	0.73	0.72	0.72
MPQA	0.71	0.71	0.71
General Inquirer	0.71	0.68	0.68
Sentiment-140	0.67	0.66	0.66
SentiWordNet 3.0	0.64	0.64	0.64

Table 11: Evaluation results - Movie reviews dataset

Lexicon	Precision	Recall	F1 Score
Lingmotif-lex	0.86	0.84	0.84
Bing Liu	0.80	0.80	0.80
Sentiment-140	0.85	0.75	0.78
SentiWordNet	0.91	0.71	0.78
General Inquirer	0.89	0.73	0.78
MPQA	0.80	0.75	0.76

Table 12: Evaluation results - STS-Gold dataset

this resource, embedded in the Lingmotif sentiment analysis system, and the way its format and valence system has evolved over the years. Three main features characterize *Lingmotif-lex* vis-à-vis other available sentiment lexicons: careful manual curation and testing on many different texts and datasets, the strong emphasis placed on multi-word expressions, and the inclusion of a valence shifters system, that accounts for valence modification by context.

We have also provided a formal evaluation of our lexicon by testing its performance on a typical sentiment classification task of two publicly available sentiment analysis datasets: the classic movie reviews used in Pang and Lee's groundbreaking 2004 paper and the STS-Gold collection of tweets. Our evaluation results show that *Lingmotif-lex* achieves significantly better performance than the other five lexicons evaluated across both datasets. Further testing, using other datasets, might be considered, but these results already offer enough evidence as to the high quality and usefulness of our resource.

8. Acknowledgements

This research was supported by Spain's MINECO through the funding of project Lingmotif2 (FFI2016-78141-P).

9. Bibliographical References

Andreevskaia, A. and Bergler, S. (2006). Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 209–216.

Andreevskaia, A. and Bergler, S. (2007). CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 117–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aue, A. and Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2200–2204, Valletta, Malta.

Choi, Y. and Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 793–801, Stroudsburg, PA, USA. Association for Computational Linguistics.

Choi, Y., Kim, Y., and Myaeng, S.-H. (2009). Domain-specific sentiment analysis using contextual feature generation. In *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 37–44, Hong Kong, China. ACM.

Christiane Fellbaum, editor. (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, Seattle, WA, USA. ACM.

Kennedy, A. and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367, Geneva, Switzerland. Association for Computational Linguistics.

Lu, Y., Castellanos, M., Dayal, U., and Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: An optimization approach. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 347–356, New York, NY, USA. ACM.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh In-*

ternational Workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA, June.

Moreno-Ortiz, A. and Pérez-Hernández, C. (2017). Tecnolengua Lingmotif at TASS 2017: Spanish Twitter dataset classification combining wide-coverage lexical resources and text features. In *TASS 2017: Workshop on Semantic Analysis at SEPLN*, pages 35–42, Murcia, Spain. SEPLN.

Moreno-Ortiz, A., Pérez-Hernández, C., and Del-Olmo, M. (2013). Managing multiword expressions in a lexicon-based sentiment analysis system for spanish. In *Proceedings of the 9th Workshop on Multiword Expressions MWE 2013*, pages 1–10, Atlanta, Georgia, USA. The Association for Computational Linguistics.

Moreno-Ortiz, A. (2017a). Lingmotif: A user-focused sentiment analysis tool. *Procesamiento del Lenguaje Natural*, 58(0):133–140, March.

Moreno-Ortiz, A. (2017b). Lingmotif: Sentiment analysis for the digital humanities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76, Valencia, Spain, April. Association for Computational Linguistics.

Moreno-Ortiz, A. (2017c). Tecnolengua lingmotif at EmoInt-2017: A lexicon-based approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 225–232, Copenhagen, Denmark, September. Association for Computational Linguistics.

Pang, B. and Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–278, Barcelona, Spain. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November.

Polanyi, L. and Zaenen, A. (2006). Contextual valence shifters. In James G. Shanahan, et al., editors, *Computing Attitude and Affect in Text: Theory and Applications*, pages 1–10. Springer Netherlands, Dordrecht.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1119369.

Riloff, E., Patwardhan, S., and Wiebe, J. (2006). Feature subsumption for opinion analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 440–448, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1610137.

Saif, H., Fernández, M., He, Y., and Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*, volume 3, Turin, Italy.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

Taboada, M., Brooks, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Tai, Y.-J. and Kao, H.-Y. (2013). Automatic domain-specific sentiment lexicon generation with label propagation. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, IIWAS '13, pages 53:53–53:62, New York, NY, USA. ACM.

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

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yu, H., Shang, J., Hsu, M., Castellanos, M., and Han, J. (2016). Data-driven contextual valence shifter quantification for multi-theme sentiment analysis. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 939–948, New York, NY, USA. ACM.