TASS 2017: Workshop on Semantic Analysis at SEPLN, septiembre 2017, págs. 35-42

Tecnolengua Lingmotif at TASS 2017: Spanish Twitter Dataset Classification Combining Wide-coverage Lexical Resources and Text Features

Tecnolengua Lingmotif en TASS 2017: Clasificación de polaridad de tuits en español combinando recursos léxicos de amplia cobertura con rasgos textuales.

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

Abstract: In this paper we describe our participation in TASS 2017 shared task on polarity classification of Spanish tweets. For this task we built a classification model based on the Lingmotif Spanish lexicon, and combined this with a number of formal text features, both general and CMC-specific, as well as single-word keywords and n-gram keywords, achieving above-average results across all three datasets. We report the results of our experiments with different combinations of said feature sets and machine learning algorithms (logistic regression and SVM).
Keywords: sentiment analysis, twitter, polarity classification

Resumen: En este artículo describimos nuestra participación en la tarea de clasificación de polaridad de tweets en español del TASS 2017. Para esta tarea hemos desarrollado un modelo de clasificación basado en el lexicón español de Lingmotif, combinado con una serie de rasgos formales de los textos, tanto generales como específicos de la comunicación mediada por ordenador (CMC), junto con palabras y unidades fraseológicas clave, lo que nos ha permitido obtener unos resultados por encima de la media en los tres conjuntos de la prueba. Mostramos los resultados de nuestros experimentos con diferentes combinaciones de conjuntos de funciones y algoritmos de aprendizaje automático (regresión logística y SVM).
Palabras clave: análisis de sentimiento, twitter, clasificación de polaridad

1 Introduction

The use of microblogging sites in general, and Twitter in particular, has become so well - established that it is now a common source to poll user opinion and even social happiness (Abdullah et al., 2015). Its relevance as a social hub can hardly be overestimated, and it is now common for traditional media to reference Twitter trending topics as an indicator of social concerns and interests.

It is not surprising, then, that Twitter datasets are increasingly being used for sentiment analysis shared tasks. The SemEval series of shared tasks included Sentiment Analysis of English Twitter content in 2013 (Nakov et al., 2013), and included other languages in later editions. The TASS Workshop on Sentiment Analysis at SEPLN series started in 2012, and continued on a yearly basis, thus being a milestone not only for Spanish Twitter content, but for sentiment analysis in general.

The General Corpus of TASS was published for TASS 2013 (Villena Román et al., 2013), introducing aspect-based sentiment analysis, consisting of over 68,000 polarity-annotated tweets. Its creation followed certain design criteria in terms of topics (politics, football, literature, and entertainment) and users.

TASS 2017 (Martínez-Cámara et al., 2017) keeps the Spain-only General Corpus of TASS, and introduces a new international corpus of Spanish tweets, named InterTASS. The InterTASS corpus adds considerable difficulty to the tasks not only because of its multi-varietal nature, but also because, unlike the General Corpus of TASS, content has

ISSN 1613-0073

not been filtered or their users selected, which introduces many and varied decoding issues.

1.1 Classification tasks

TASS 2017 proposes two classification tasks. Task 1 focuses on sentiment analysis at the tweet level, while Task 2 deals with aspect-based sentiment classification. We took part in Task 1, since we have not yet tackled aspect-based sentiment analysis. The aim of this task is the automatic classification of tweets in one of 4 levels: POSITIVE, NEGATIVE, NEUTRAL, and NONE.

The NEUTRAL/NONE distinction introduces added difficulty to the classification task. Tweets annotated as NONE are supposed to express no sentiment whatsoever, as in informative or declarative texts, whereas the NEUTRAL category of tweets is meant to qualify tweets where both positive and negative opinion is expressed, but they cancel each other out, resulting in a neutral overall message.

We believe this distinction is too fuzzy to be annotated reliably. First, precise balance of polarity is hardly ever found in any message where sentiment is expressed: the message is usually "negative/positive situation *x*, *somehow* counterbalanced by positive/negative situation *y*", with an entailment that the result is tilted to either side. The following are examples of tweets tagged as NEUTRAL in the training set:

- 768547351443169284 Parece que las cosas no te van muy bien, espero que todo mejore, que todo el mundo merece ser feliz.
- 770417499317895168 No hay nada más bonito q separarse d una persona y q al tiempo t diga q t echa de menos... pero a mi no m va a pasar

We also found a number of examples where tweets that clearly fell into NONE cases, where wrongly annotated as NEUTRAL:

- 768588061496209408 Estas palabras, del Poema, INSTANTES, son de Nadine Stair. Escritora norteamericana, a la q le gustan los helados.
- 767846757996847104 pues imaginate en una casa muy grande
- 769993102442524674 Ninguno de los clubes lo hizo oficial pero se dice que sí

These annotation issues are to be expected, due to the added cognitive load that is placed on the annotators, as other researchers have pointed out (Mohammad and Bravo-Marquez, 2017a). Also, its presence makes it more difficult to compare results with those of other sentiment classification shared tasks, where the NONE class is not considered.

1.2 Lexicon-based Sentiment Analysis

Within Sentiment Analysis it is common to distinguish corpus-based approaches from lexicon-based approaches. Although a combination of both methods can be found in the literature (Riloff, Patwardhan, and Wiebe, 2006), Lexicon-based approaches are usually preferred for sentence-level classification (Andreevskaia and Bergler, 2007), whereas corpus-based, statistical approaches are preferred for document-level classification.

Using sentiment dictionaries has a long tradition in the field. WordNet (Fellbaum, 1998) has been a recurrent source of lexical information (Kim and Hovy, 2004; Hu and Liu, 2004; Adreevskaia 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, Wiebe, and Hoffmann, 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).

However, the valence of a given word may vary greatly from one domain to another, a fact well recognized in the literature (Aue and Gamon, 2005; Pang and Lee, 2008; Choi, Kim, and Myaeng, 2009), which causes problems when a sentiment lexicon is the only source of knowledge. A number of solutions have been proposed, mostly using ad hoc dic-

[1] The use of the terms *valence* and *polarity* is 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.

tionaries, sometimes created automatically from a domain-specific corpus (Tai and Kao, 2013; Lu et al., 2011).

Our approach to using a lexicon takes some ideas from the aforementioned approaches. We describe it in the next section.

2 System description

Our system for this polarity classification task relies on the availability of rich sets of lexical, sentiment, and (formal) text features, rather than on highly sophisticated algorithms. We basically used a logistic regression classifier trained on the optimal set of features after many feature combinations were tried on the training set. We also tried a SVM classifier on the same feature sets, but we consistently obtained poorer results compared to the logistic regression classifier. Parameter finetuning on each classifier was very limited; we simply performed a grid search on the C parameter, which threw 100 as optimal. For the SVM classifier we found the RBF kernel to perform better than the linear kernel[2]. We mostly focused on feature selection and combination.

We obtained good results on the three test datasets, with some important differences between the InterTASS and General datasets. Results, however, were not as good as we had anticipated based on our experiments on the training datasets. We discuss this in section 3 below. Here we describe our general system architecture and feature sets.

This TASS shared task is our first experience with Twitter data sentiment classification proper, although we had the related experience from our recent participation in WASSA-2017 Shared Task on Emotion Intensity (Mohammad and Bravo-Marquez, 2017b). From this shared task we learnt the relevance and impact that other, non-lexical text features can have in microblogging texts.

Since our focus was on identifying the predictive power of classification features, and intended to perform many experiments with features combinations, we designed a simple tool to facilitate this.

This tool, Lingmotif Learn, is a GUI-enabled convenience tool that manages datasets and uses the Python-based scikit-learn (Pedregosa et al., 2011) machine learning toolkit. It facilitates loading and preprocessing datasets, getting the text run trough the Lingmotif SA engine, and feeding the resulting data into one of several machine learning algorithms. Lingmotif Learn is able to extract both Sentiment features and non-sentiment features, such as raw text metrics and keywords, and it makes it easy to experiment with different feature set combinations.

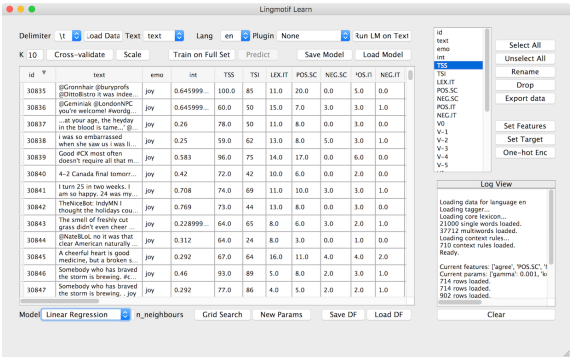

Figure 1: Lingmotif Learn

2.1 The Lingmotif tool

Sentiment features are returned by the Lingmotif SA engine. Lingmotif (Moreno-Ortiz, 2017a) is a user-friendly, multilingual, text analysis application with a focus on sentiment analysis that offers several modes of text analysis. It is not specifically geared towards any particular type of text or domain. It can analyze long documents, such as narratives, medium-sized ones, such as political speeches and debates, and short to very short texts, such as user reviews and tweets. For each of these, the tool offers different outputs and metrics.

For large collections of short texts, such as Twitter datasets, it provides a multi-document mode whose default output is classification. In the current publicly available version this classification is entirely based on the Text Sentiment Score (TSS), which attempts to summarize the text's overall polarity on a 0-100 scale. TSS is calculated as a function of the text's positive and negative scores and the sentiment intensity, which reflects the proportion of sentiment to non-sentiment lexical items in the text. Specific details on TSS calculation can be found in Moreno-Ortiz (2017a). A description of its applications is found in Moreno-Ortiz (2017b).

Lingmotif results are generated as a HTML/Javascript document, which is saved

[2]For the RBF kernel we used *gamma=0.001*, *C=100*. For the linear kernel we used *C=1000*.

Name	Description
TSS	Text Sentiment Score
TSI	Text Sentiment Intensity
SENT.IT	Number of lexical Items
POS.SC	Positive score
NEG.SC	Negative score
POS.IT	Number of positive items
NEG.IT	Number of negative items
NEU.IT	Number of neutral items
SPLIT1.TSS	TSS for split 1 of text
SPLIT2.TSS	TSS for split 2 of text
SENTENCES	Number of sentences
SHIFTERS	Number of sentiment shifters

Table 1: Sentiment feature set

locally to a predefined location and automatically sent to the user's default browser for immediate display. Internally, the application generates results as an XML document containing all the relevant data; this XML document is then parsed against one of several available XSL templates, and transformed into the final HTML.

Lingmotif Learn simply plugs into the internally generated XML document to retrieve the desired sentiment analysis data, and appends the data to each tweet as features.

2.2 Sentiment features

Table 1 summarizes the sentiment-related feature set generated by the Lingmotif engine.

Most of these features are included in the original Lingmotif engine, but for this occasion we experimented with text splits to test the relevance of the position of the sentiment words in the tweet. The features SPLIT1.TSS and SPLIT2.TSS are the combined sentiment score for each half of the tweet. The assumption was that sentiment words used towards the end of the tweet may have more weight on the overall tweet polarity. This might be helpful especially for the P/N/NEU distinction. Neutral tweets are supposed to have some balance between positivity and negativity. In our tests with the training set, however, adding these features did not improve results. We also experimented with 3 splits, with the same results. These features were thus discarded for test set classification.

Some of these features are in fact redundant. Notably, TSS already encapsulates POS.SC, NEG.SC, and NEU.IT. In our tests, the classifier performed better using just the POS.SC and NEG.SC values, than our calculated TSS, so we only used these two features.

Name	Description
SENTENCES	Number of sentences
TT.RATIO	Type/Token ratio
LEX.ITEMS	Number of lexical items
GRAM.ITEMS	Number of grammatical items
VB.ITEMS	Number of verbs
NN.ITEMS	Number of nouns
NNP.ITEMS	Number of proper nouns
JJ.ITEMS	Number of adjectives
RB.ITEMS	Number of adverbs
CHARS	Number of characters
INTENSIFIERS	Number of intensifiers
CONTRASTERS	Number of contrast words
EMOTICONS	Number of emoticons/emojis
ALL.CAPS	Number of upper case words
CHAR.NGRAMS	Number of character ngrams
X.MARKS	Number of exclamation marks
Q.MARKS	Number of question marks
QUOTE.MARKS	Number of quotation marks
SUSP.MARKS	Number of suspension marks
X.MARKS.SEQS	Number of x.marks sequences
Q.MARKS.SEQS	Number of q.marks sequences
XQ.MARKS.SEQS	Number of x/q marks sequences
HANDLES	Number of Twitter handles
HASHTAGS	Number of hashtags
URLS	Number of URL's

Table 2: Text feature set

2.3 Text features

Raw text features are commonly used in sentiment analysis shared tasks successfully (e.g. Mohammad, Kiritchenko, and Zhu (2013), Kiritchenko et al. (2014)), including previous editions of TASS (Cerón-Guzmán, 2016). The role of some of them is rather obvious; the presence of emoticons or exclamation marks, for example, usually determines (strong) sentiment or opinion, thus being a good candidate predictor for the NONE vs rest distinction. The role of others, however, is not as clear. For example, we consistently obtained better results using the GRAM.ITEMS feature, whereas the number of lexical items was not a good predictor. The number of verbs, adjectives and adverbs also proved to be useful, whereas the number of nouns did not.

Table 2 contains the full list of text features we experimented with.

2.4 Keyword features

In order to account for words and expressions that convey sentiment but may not be included in the sentiment lexicon, we experimented with automatic keyword extraction for each of the classes in the training set. Automatic keyword and keyphrase extraction is a well developed field and a number of tools and methodologies have been pro-

Name	Description
P.KW	Positive keywords
P.NG.KW	Positive ngram keywords
P.HANDLES	Positive handles
N.KW	Negative keywords
N.NG.KW	Negative ngram keywords
N.HANDLES	Negative handles
NEU.KW	Neutral keywords
NEU.NG.KW	Neutral ngram keywords
NEU.HANDLES	Neutral handles
NONE.KW	None keywords
NONE.NG.KW	None ngram keywords
NONE.HANDLES	None handles

Table 3: Keywords feature set

posed. Hasan and Ng (2014) provide a good overview of the state-of-the-art techniques for keyphrase extraction.

We used a very simple approach that consisted in comparing frequencies of single words and ngrams (2 to 4 words) on a one-vs-rest basis for each of our four classes, for words and ngrams with a minimum frequency of 2. We calculated and ranked keyness based on the chi-square statistic, and then manually removed irrelevant results. We ended up with a list of 100 keywords and 100 keyphrases for each class. We did the same for Twitter handles.

Using the keywords feature set improved results considerably in our tests with the training set. However, this improvement did not transfer well to the test sets, especially in the case of the InterTASS dataset. We further discuss this issue in section 3.

3 Experiments and Results

Tables 4, 5, and 6 show our results for each of the test sets. Although performance is strong across all three, there clearly is a difference between the General TASS datasets, on the one hand, and the InterTASS dataset on the other.

Experiment	Macro-F1	Accuracy
sent-only	0.456	0.582
run3	0.441	0.576
sent-only-fixed	0.441	0.595

Table 4: Official results for the InterTASS test set

We believe this is due to two main reasons. First, the General training set (7,218 tweets) is much larger than the InterTASS training set (1,514 tweets, using both the training and development datasets). This of course provides a much more solid training base for

Experiment	Macro-F1	Accuracy
run3	0.528	0.657
final	0.517	0.632
no_ngrams	0.508	0.652

Table 5: Official results for the General TASS test set

Experiment	Macro-F1	Accuracy
run3	0.521	0.638
final	0.488	0.618
run4	0.483	0.612

Table 6: Official results for the General TASS-1k test set

the former than the latter. All our models were trained on one dataset where both training datasets (General and InterTASS) where merged. Perhaps better results would have been obtained by training on each dataset separately.

The other reason for poorer performance on the InterTASS test set concerns the very different nature of the datasets. The General Corpus of TASS consists of tweets generated by public figures (artists, politicians, journalists) with a large number of followers. Such Twitter users are more predictable both in terms of the content of their tweets and the language they use. They are also Castilian Spanish speakers entirely. Most of these tweets contain very compact but carefully chosen language, expressing users' opinion or evaluation of polically or socially relevant events. On the other hand, the interTASS corpus shows much more variability; first, the tweets were collected not only from Spain, but from several Latin American countries, which introduces important lexical variability. Second, no user selection is apparent. Tweets were randomly collected from the whole Spanish speaking user base. This introduces spelling errors and a much more colloquial and chatty language. Non-lexical linguistic features, such as exclamation marks, emojis or emoticons, are recurrent, as are, user-to user messages, which are of course hard-to-decode, since they presuppose certain privately shared knowledge. These issues have obviously affected the performance of all TASS participants, as is clear from the final leader board.

We obtained the best results for the General datasets with our *run3* experiment, where we combined a selection of features from the three feature sets listed in tables

Features	
POS.SC	NEU.KW
NEG.SC	NEU.NG.KW
VB.ITEMS	NEU.HANDLES
JJ.ITEMS	NONE.KW
RB.ITEMS	NONE.NG.KW
GRAM.ITEMS	NONE.HANDLES
N.CHARS	EMOTICONS
INTENSIFIERS	ALL.CAPS
CONTRASTERS	CHAR.NGRAMS
P.KW	X.MARKS
P.NG.KW	Q.MARKS
P.HANDLES	SUSP.MARKS
N.KW	HASHTAGS
N.NG.KW	HANDLES
N.HANDLES	URLS

Table 7: *run3* experiment feature set

Features	
POS.SC	HANDLES
NEG.SC	EMOTICONS
VB.ITEMS	ALL.CAPS
JJ.ITEMS	CHAR.NGRAMS
RB.ITEMS	X.MARKS
GRAM.ITEMS	Q.MARKS
N.CHARS	SUSP.MARKS
INTENSIFIERS	URLS
CONTRASTERS	HASHTAGS

Table 8: *sent-only* experiment feature set

1, 2, and 3. This selection was in fact the optimal we found during our cross-validation tests on the training dataset. Table 7 lists the feature set used in this experiment.

Concerning the InterTass test set, the best results were obtained with the *sent-only* experiment, where a reduced set of features was used. We list these features in table 8.

We obtained better results for the InterTASS test set using this reduced set of features because the keyword sets were causing noise, since they were extracted using the whole training set, which contained a much larger proportion of tweets from the General TASS dataset.

Another important aspect is the large difference that we encountered between our own tests on the training datasets and our final (official) results. For the General corpus of TASS, we consistently obtained very high F1 scores (upwards of 0.73) using the keyword set, but much closer to the official results without them. This is a clear indication of model overfitting, with an obvious negative impact on the classification of the test set. After this became apparent on our first results upload, we corrected by reducing the sets of keywords, keyphrases and user handles, which resulted in better overall results.

4 Conclusions

This shared task has served us to assess the usefulness of many different features as predictors of polarity classification in Spanish tweets. The differing sizes and characteristics of the training and test datasets determined to some extent our results, but we also felt we overfitted our model with too large a selection of keywords, which threw overoptimistic results in our tests.

Our results on par with other participants who used more sophisticated systems from the technical perspective, which is also an indication of the salient role that curated, high-quality lexical resources play in sentiment analysis.

We also experienced the negative impact of model overfitting and learnt how to limit its effects. We plan to use this knowledge in future versions of Lingmotif, which currently uses sentiment features exclusively. It is obvious that combining those with other formal features can improve results considerably.

Acknowledgments

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

References

Abdullah, S., E. L. Murnane, J. M. Costa, and T. Choudhury. 2015. Collective smile: Measuring societal happiness from geolocated images. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 361–374, New York, NY, USA. ACM.

Adreevskaia, A. and S. Bergler. 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 S. Bergler. 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 M. Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. Borovets, Bulgaria.

Cerón-Guzmán, J. A. 2016. Jacerong at tass 2016: An ensemble classifier for sentiment analysis of spanish tweets at global level. In *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with 32nd SEPLN Conference (SEPLN 2016)*, pages 35–39, Salamanca, Spain. SEPLN.

Choi, Y. and C. Cardie. 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.

Choi, Y., Y. Kim, and S.-H. Myaeng. 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.

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

Hasan, K. S. and V. Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273.

Hatzivassiloglou, V. and K. R. McKeown. 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 B. Liu. 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 D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

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

Kiritchenko, S., X. Zhu, C. Cherry, and S. Mohammad. 2014. Nrc–canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Lu, Y., M. Castellanos, U. Dayal, and C. Zhai. 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.

Martínez-Cámara, E., M. C. Díaz-Galiano, M. Á. García-Cumbreras, M. García-Vega, and J. Villena-Román. 2017. Overview of tass 2017. In J. Villena Román, M. Á. García Cumbreras, E. Martínez-Cámara, M. C. Díaz Galiano, and M. García Vega, editors, *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volume 1896 of *CEUR Workshop Proceedings*, Murcia, Spain, September. CEUR-WS.

Mohammad, S. and F. Bravo-Marquez. 2017a. Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada.

Mohammad, S. and F. Bravo-Marquez. 2017b. Wassa-2017 shared task on emotion intensity. In *Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media*, Copenhagen, Denmark, September.

Mohammad, S. M., S. Kiritchenko, and X. Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

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.

Nakov, P., Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, June.

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

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

Polanyi, L. and A. Zaenen. 2006. Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*. Springer, Dordrecht, The Netherlands, shanahan, james g., qu, yan, wiebe, janyce edition, pages 1–10.

Riloff, E., S. Patwardhan, and J. Wiebe. 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.

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

Tai, Y.-J. and H.-Y. Kao. 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.

Villena Román, J., S. Lana Serrano, E. Martínez Cámara, and J. C. González Cristóbal. 2013. Tass - workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50:37–44.

Wilson, T., J. Wiebe, and P. Hoffmann. 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.