UWB at SemEval-2018 Task 1: Emotion Intensity Detection in Tweets

Pavel Přibáň[1,2], **Tomáš Hercig**[1,2], and **Ladislav Lenc**[1]

[1]NTIS – New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
[2]Department of Computer Science and Engineering,
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{pribanp,tigi,llenc}@kiv.zcu.cz
http://nlp.kiv.zcu.cz

Abstract

This paper describes our system created for the SemEval-2018 Task 1: Affect in Tweets (AIT-2018). We participated in both the regression and the ordinal classification subtasks for emotion intensity detection in English, Arabic, and Spanish.

For the regression subtask we use the AffectiveTweets system with added features using various word embeddings, lexicons, and LDA. For the ordinal classification we additionally use our Brainy system with features using parse tree, POS tags, and morphological features. The most beneficial features apart from word and character n-grams include word embeddings, POS count and morphological features.

1 Introduction

The task of Detecting Emotion Intensity assigns the intensity to a tweet with given emotion. The emotions include anger, fear, joy, and sadness. The intensity is either on a scale of zero to one for the regression subtask, or one of four classes (0:no, 1: low, 2: moderate, 3: high) for the classification subtask. The task was prepared in three languages: English, Arabic, and Spanish. For each language there are four training and test sets of data – one for each emotion. The data creation is described in (Mohammad and Kiritchenko, 2018) and detailed description of the task is in (Mohammad et al., 2018).

We participated in the emotion intensity regression task (**EI-reg**) and in the emotion intensity ordinal classification task (**EI-oc**) in English, Arabic and Spanish.

2 System Description

We used two separate systems for ordinal classification – AffectiveTweets (Section 3) and Brainy (Section 4). For the regression task we just use the AffectiveTweets system. We train a separate model for each emotion. The Brainy system performed better in our pre-evaluation experiments on the development data for all emotions in Spanish and for fear and joy emotions in Arabic.

3 AffectiveTweets System

3.1 Tweets Preprocessing

Tweets often contain slang expressions, misspelled words, emoticons or abbreviations and it's needed to make some preprocessing steps before extracting features. First, every tweet was tokenized using *TweetNLP*[1](Gimpel et al., 2011). Then the AffectiveTweets[2] (Mohammad and Bravo-Marquez, 2017) package for Weka machine learning workbench (Hall et al., 2009) was used for feature extraction. The following steps were applied on tokens for every language in both tasks:

1. Tokens were converted to lowercase
2. URL links were replaced with *http://www.url.com* token
3. Twitter usernames (tokens starting with @) were replaced with @*user* token
4. Tokens containing sequences of letters occurring more than two times in a row were replaced with two occurrences of them (e.g. *huuuungry* is reduced to *huungry*, *looooove* to *loove*)
5. Common sequences of words and emojis were divided by space (e.g. token „*nice:D:D*“ was divided into two tokens „*nice*“ and „*:D:D*“)

[1]http://www.cs.cmu.edu/~ark/TweetNLP/
[2]https://affectivetweets.cms.waikato.ac.nz/

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 133–140
New Orleans, Louisiana, June 5–6, 2018.

These steps lead to reduction of feature space as shown in (Go et al., 2009). We also used some individual preprocessing for Arabic language. After the above described steps every token was also processed via *Stanford Word Segmenter*[3](Monroe et al., 2014). When using word embeddings, we transformed Arabic words from regular UTF-8 Arabic to a more ambiguous form[4]. This was done only for word embedding features.

3.2 Features

Our AffectiveTweets system used combinations of features that are described in this section. The submitted combination of features is shown in Table 1.

- **Word n-grams ($\mathbf{WN}_i^n$):** word n-grams[5]from i to n (for $i = 1$, $n = 2$, *unigrams* and *bigrams* were used).
- **Character n-grams ($\mathbf{ChN}_i^n$):** character n-grams[5] from i to n (for $i = 2$, $n = 3$ character *bigrams* and *trigrams* were used).
- **Word Embeddings (WE):** an average of the word embeddings of all the words in a tweet.
- **Affective Lexicons (L):** we used AffectiveTweets package to extract features from affective lexicons. In every language we also used SentiStrength **(L-se)** lexcion-based method (Thelwall et al., 2012).
- **LDA – Latent Dirichlet Allocation ($\mathbf{D}_n$):** topic distribution of tweet, that is obtained from our pre-trained model, n indicates number of topics in model (for $n = 5$, feature vector with dimension 5 will be produced and each component of the vector refers to one topic). We used LDA features only in AffectiveTweets system.

3.2.1 English Word Embeddings:

- **Ultradense Word Embeddings (WE-ue):** Rothe et al. (2016) created embeddings in the Twitter domain.
- **Baseline Word Embeddings (WE-b):** Mohammad and Bravo-Marquez (2017) created embeddings from the Edinburgh Twitter Corpus (Petrović et al., 2010).

3.2.2 Spanish Word Embeddings:

- **Ultradense Word Embeddings (WE-us):** Rothe et al. (2016) created embeddings from web domain.
- **FastText Word Embeddings (WE-ft):** Bojanowski et al. (2016) trained embeddings on Wikipedia.

3.2.3 Arabic Word Embeddings:

- **Zahran et al. (2015) Word Embeddings (var-SG**, **var-GloVe**, and **var-CBOW)**
- **Soliman et al. (2017) Word Embeddings (tw-SG**, **tw-CBOW**, **web-SG**, **web-CBOW**, **wiki-SG**, and **wiki-CBOW)**

Mentioned Arabic word embeddings were created with Global Vectors (GloVe) (Pennington et al., 2014) and Word2Vec toolkit (Mikolov et al., 2013) using skip-gram (SG) model and continuous bag-of-words (CBOW) model. These Arabic word embeddings were trained on different data domains – Twitter (tw), web pages (web), Wikipedia (wiki), and their combination (var) for more details see the cited papers.

3.2.4 English lexicons (L-en):

– We used all affective lexicons from the AffectiveTweets package.

3.2.5 Spanish lexicons (L-es):

– Translated NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013)

– Emotion Lexicon (Sidorov et al., 2012)

– Polarity lexicon (Urizar and Roncal, 2013)

– Expanded Word-Emotion Association Lexicon (Bravo-Marquez et al., 2016) (we translated this lexicon to Spanish)

– iSOL (Molina-González et al., 2013)

– ML-SentiCon (Cruz et al., 2014)

– Ultradense lexicon (Rothe et al., 2016)

– LYSA Twitter lexicon (Vilares et al., 2014)

3.2.6 Arabic lexicons (L-ar):

– Translated NRC Word-Emotion Association Lexicon

– Translation of Bing Liu's Lexicon

– Arabic Emoticon Lexicon

– Arabic Hashtag Lexicon

[3] https://nlp.stanford.edu/software/segmenter.shtml

[4] Some characters were replaced, for more details see (Soliman et al., 2017).

[5] Value of each feature is set to its frequency in the tweet

	English	Arabic	Spanish
Regression			
anger	L-en, D_{500}	var-SG, L-ar, D_{250}, WN^1_1	L-en, L-es, WE-us, WN^1_1
fear	L-en, L-se, WE-b	var-SG, L-ar, D_{250}	L-en, L-es, L-se, WE-us, WN^1_1, D_{1000}
joy	L-en, L-se, WE-b	var-SG, L-ar D_{250}	L-es, WE-us, WN^2_1, ChN^3_2
sadness	L-en, L-se, WE-b	var-SG, L-ar	L-en, L-es, L-se, WE-us, WN^2_1, D_{1000}
Classification			
anger	L-en, D_{250}	WN^1_1	
fear	L-en, L-se, WE-b, WN^2_1, D_{250}		
joy	L-en, L-se, WE-b, WN^2_1		
sadness	L-en, L-se, D_{250}	var-CBOW, L-ar, L-se, WN^1_1, D_{250}	

Table 1: Used features in the AffectiveTweets system

– Arabic Hashtag Lexicon (dialectal)

– Translated NRC Hashtag Sentiment Lexicon

– SemEval-2016 Arabic Twitter Lexicon

Lexicons are described in (Mohammad and Turney, 2013; Mohammad et al., 2016a; Salameh et al., 2015; Mohammad et al., 2016b).

3.3 Model Training

In our AffectiveTweets system we used an L_2-regularized L_2-loss SVM regression and classification model with the regularization parameter C set to 1, implemented in LIBLINEAR Library (Fan et al., 2008)[6].

3.4 LDA Training

To use topics created with LDA (Latent Dirichlet Allocation) (Blei et al., 2003) as features, we trained our own models for every language. Tweets used to train the Arabic and Spanish models were taken from SemEval-2018 AIT DISC corpus (Mohammad et al., 2018) and tweets for English model were taken from Sentiment140[7] training data (Go et al., 2009). We trained our LDA models with LDA implementation from MALLET[8](McCallum, 2002).

We used the same preprocessing for LDA as for regular feature extraction. Additionally we removed stopwords and following special characters [, . ! -]. Tokens from Spanish tweets were stemmed with *Snowball*[9] stemming algorithm.

4 Brainy System

We use Maximum Entropy classifier from Brainy machine learning library (Konkol, 2014) and UDPipe (Straka et al., 2016) for preprocessing and doesn't use any lexicons, just word embeddings. The system is based on (Hercig et al., 2016).

4.1 Preprocessing

The same preprocessing has been done for all datasets. We use UDPipe (Straka et al., 2016) with Spanish Universal Dependencies 1.2 models and Arabic Universal Dependencies 2.0 models for POS tagging and lemmatization. Tokenization has been done by TweetNLP tokenizer (Owoputi et al., 2013). We further replace all user mentions with the token "@USER" and all links with the token "$LINK".

4.2 Features

The Brainy system used the following features. The exact combination of features for each emotion and the change in performance caused by its removal is shown in Table 9.

- **Character n-grams (ChN_n):** Separate binary feature for each character n-gram in the utterance text. We do it separately for different orders $n \in \{1, 2, 3, 4, 5\}$ and remove n-grams with frequency t.

- **Bag of Words (BoW):** We used bag-of-words representation of a tweet, i.e. separate binary feature representing the occurrence of a word in the tweet.

- **Bag of Morphological features (BoM):** for all verbs in the tweet. The morphological features[10] include abbreviation, aspect, definiteness, degree of comparison, evidentiality, mood, polarity, politeness, possessive, pronominal type, tense, verb form, and voice.

[6] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[7] http://help.sentiment140.com/

[8] http://mallet.cs.umass.edu/

[9] http://snowballstem.org/

[10] http://universaldependencies.org/u/feat/index.html

- **Bag of POS (BoPOS):** We used bag-of-words representation of a tweet, i.e. separate binary feature representing the occurrence of a POS tag in the tweet.

- **Bag of Parse Tree Tags (BoT):** We used bag-of-words representation of a tweet, i.e. separate binary feature representing the occurrence of a parse tree tag in the tweet. We remove tags with a frequency ≤ 2.

- **Emoticons (E):** We used a list of positive and negative emoticons (Montejo-Ráez et al., 2012). The feature captures the presence of an emoticon within the text.

- **First Words (FW):** Bag of first five words with at least 2 occurrences.

- **Last Words (LW):** Bag of last five words with at least 2 occurrences.

- **Last BoM (LBoM):** Bag of last five morphological features (see BoM) with at least 2 occurrences.

- **FastText (FT):** An average of the FastText (Bojanowski et al., 2016) word embeddings of all the words in a tweet.

- **N-gram Shape (NSh):** The occurrence of word shape n-gram in the tweet. Word shape assigns words into one of 24 classes[11] similar to the function specified in (Bikel et al., 1997). We consider unigrams, bigrams, and trigrams with frequency ≤ 2.

- **POS Count Bins (POS-B):** We map the frequency of POS tags in a tweet into a one-hot vector with length three and use this vector as binary features for the classifier. The frequency belongs to one of three equal-frequency bins[12]. Each bin corresponds to a position in the vector. We remove POS tags with frequency $t \leq 5$.

- **TF-IDF:** Term frequency – inverse document frequency of a word computed from the training data for words with at least 5 occurrences and at most 50 occurrences.

[11] We use edu.stanford.nlp.process.WordShapeClassifier with the WORDSHAPECHRIS1 setting available in Stanford CoreNLP library (Manning et al., 2014).

[12] The frequencies from the training data are split into three equal-size bins according to 33% quantiles.

Emotion intensity regression – Pearson (all instances)					
embeddings	avg	anger	fear	joy	sadness
var-SG	**0.564**	0.505	**0.569**	0.577	0.605
var-GloVe	0.523	0.489	0.520	0.529	0.557
var-CBOW	0.557	0.492	0.557	0.555	**0.622**
tw-SG	0.541	**0.513**	0.520	**0.580**	0.552
tw-CBOW	0.447	0.413	0.424	0.472	0.478
web-SG	0.492	0.419	0.465	0.559	0.526
web-CBOW	0.410	0.339	0.423	0.466	0.411
wiki-SG	0.440	0.345	0.443	0.505	0.469
wiki-CBOW	0.291	0.281	0.244	0.315	0.322
Emotion intensity classification – Pearson (all classes)					
var-SG	0.386	0.430	0.387	0.471	0.418
var-GloVe	0.318	0.410	0.383	0.430	0.385
var-CBOW	**0.397**	0.451	**0.496**	**0.536**	**0.470**
tw-SG	0.360	**0.480**	0.386	0.439	0.416
tw-CBOW	0.338	0.368	0.301	0.369	0.344
web-SG	0.325	0.426	0.424	0.375	0.388
web-CBOW	0.190	0.314	0.317	0.269	0.273
wiki-SG	0.244	0.396	0.368	0.370	0.345
wiki-CBOW	0.275	0.252	0.284	0.293	0.276

Table 2: Arabic embeddings experiments results

Emotion intensity regression – Pearson (all instances)					
embeddings	avg	anger	fear	joy	sadness
WE-us	**0.559**	**0.464**	**0.581**	**0.581**	**0.611**
WE-ft	0.510	0.369	0.577	0.528	0.565
Emotion intensity classification – Pearson (all classes)					
WE-us	**0.429**	**0.422**	0.382	0.478	0.434
WE-ft	0.407	0.256	**0.428**	**0.481**	**0.462**

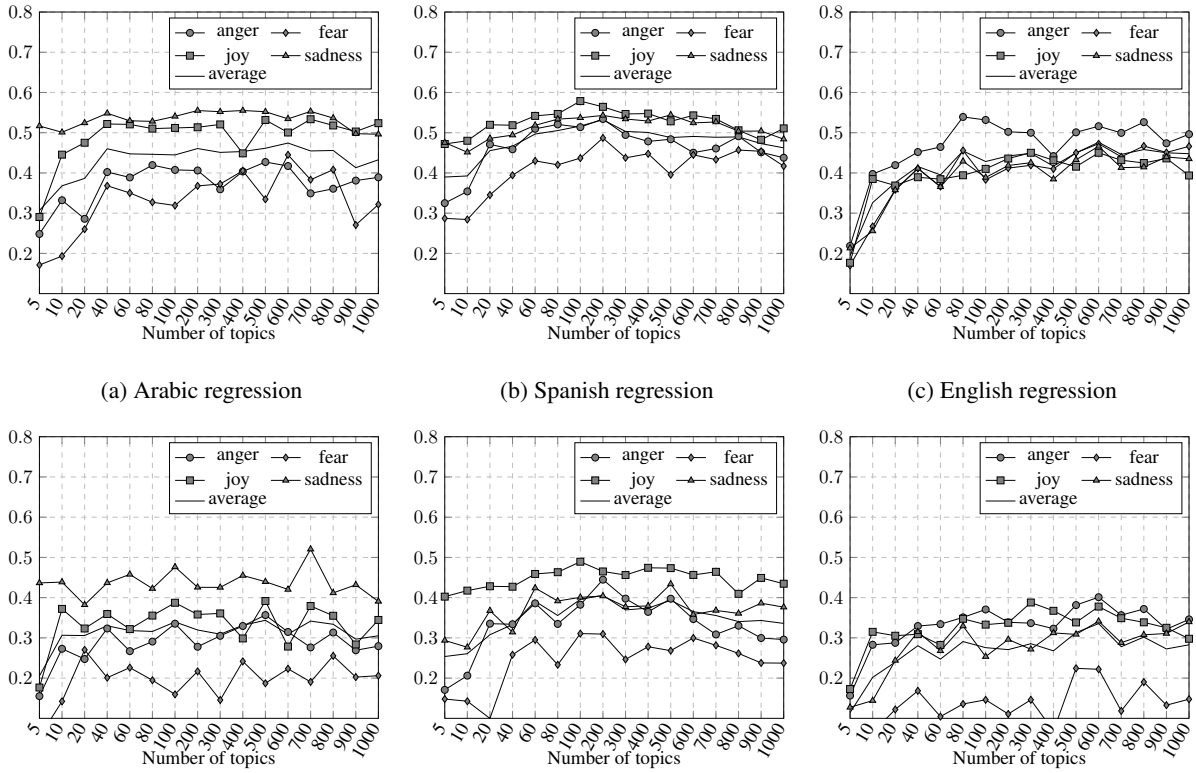
Table 3: Spanish embeddings experiments results

Emotion intensity regression – Pearson (all instances)					
embeddings	avg	anger	fear	joy	sadness
WE-ue	**0.598**	**0.594**	**0.595**	**0.586**	**0.593**
WE-b	0.541	0.475	0.549	0.456	0.505
Emotion intensity classification – Pearson (all classes)					
WE-ue	**0.479**	**0.412**	**0.507**	**0.438**	**0.459**
WE-b	0.456	0.212	0.499	0.336	0.376

Table 4: English embeddings experiments results

- **Text Length Bins (TL-B):** We map the tweet length into a one-hot vector with length three and use this vector as binary features for the classifier. The length of a tweet belongs to one of three equal-frequency bins[12]. Each bin corresponds to a position in the vector.

- **Verb Bag of Words (V-BoW):** Bag of words for parent, siblings, and children of the verb from the sentence parse tree.



(a) Arabic regression (b) Spanish regression (c) English regression

(d) Arabic classification (e) Spanish classification (f) English classification

Figure 1: LDA performance based on number of topics, the y-axis denotes Pearson correlation

5 Experiments

All presented experiments are evaluated on the test data for the given task.

We performed ablation experiments to see which features are the most beneficial (see Table 9, 8, and 10). Numbers represent the performance change when the given feature is removed[13].

Word embeddings features have a great impact on system performance, so we compared several word embeddings for every language (Table 2, 3, and 4). For English was best **WE-ue** word embeddings, but for submission we used **WE-b** word embeddings, because it worked better on dev data. In Spanish tweets the **WE-us** word embeddings outperformed the **WE-ft** word embeddings in regression and **WE-us** was better for classification in anger and on average of all emotions. For classification in Arabic was **var-CBOW** best on every emotion except anger and for regression **var-SG** worked best on average and on fear.

We also experimented with only LDA features to find out how the numbers of topics in LDA model affect the performance (see Figure 1). We started with models containing 5 topics and continued up to 1000 (step was non-equidistantly increased). Our experiments suggest that the best setting is around 200-300 topics. We selected the number of topics based on the performance on the development data.

6 Results

Our results in the emotion intensity regression subtask are in Table 5 and our results in the emotion intensity ordinal classification subtask are in Table 6 and Table 7. The system settings and features for each language and emotion were selected based on our pre-evaluation experiments with evaluation on the development data.

7 Conclusion

We competed in the emotion intensity regression and ordinal classification tasks in English, Arabic and Spanish.

Our ranks are 27th out of 48 for English, 5th out of 14 for Arabic, and 5th out of 16 for Spanish for the regression task and 21st out of 39 for English, 5th out of 14 for Arabic, and 5th out of 16 for Spanish for the ordinal classification task.

[13]The lowest number denotes the most beneficial feature

Subtask	System	**Pearson** (all instances) macro-avg	anger	fear	joy	sadness	**Pearson** (gold in 0.5 – 1) macro-avg	anger	fear	joy	sadness
EI-reg-EN	AffectiveTweets	0.642 (27)	0.640 (27)	0.642 (27)	0.652 (24)	0.636 (23)	0.478 (25)	0.503 (29)	0.433 (27)	0.457 (23)	0.517 (23)
EI-reg-AR	AffectiveTweets	0.574 (5)	0.487 (6)	0.559 (5)	0.619 (6)	0.631 (5)	0.417 (6)	0.332 (6)	0.485 (3)	0.327 (7)	0.523 (4)
EI-reg-ES	AffectiveTweets	0.630 (5)	0.542 (5)	0.688 (3)	0.646 (5)	0.644 (4)	0.496 (3)	0.435 (2)	0.517 (3)	0.527 (3)	0.507 (4)

Table 5: Pearson correlation for the emotion intensity regression task

Subtask	System	**Pearson** (all classes) macro-avg	anger	fear	joy	sadness	**Pearson** (some-emotion) macro-avg	anger	fear	joy	sadness
EI-oc-EN	AffectiveTweets	0.506 (21)	0.477 (23)	0.470 (17)	0.555 (19)	0.522 (22)	0.346 (23)	0.308 (25)	0.273 (21)	0.452 (21)	0.350 (25)
EI-oc-AR	AT&Brainy	0.394 (5)	0.327 (5)	0.345 (5)	0.437 (5)	0.467 (5)	0.280 (5)	0.246 (6)	0.246 (6)	0.351 (5)	0.277 (7)
EI-oc-ES	Brainy	0.504 (5)	0.361 (7)	0.606 (3)	0.544 (5)	0.506 (5)	0.410 (5)	0.267 (6)	0.499 (2)	0.420 (6)	0.452 (5)

Table 6: Pearson correlation for the emotion intensity ordinal classification task

Subtask	System	**Kappa** (all classes) macro-avg	anger	fear	joy	sadness	**Kappa** (some-emotion) macro-avg	anger	fear	joy	sadness
EI-oc-EN	AffectiveTweets	0.494 (21)	0.467 (19)	0.450 (14)	0.548 (17)	0.510 (19)	0.290 (23)	0.269 (23)	0.166 (20)	0.420 (20)	0.303 (24)
EI-oc-AR	AT&Brainy	0.386 (5)	0.324 (5)	0.327 (5)	0.428 (5)	0.464 (5)	0.241 (5)	0.219 (5)	0.178 (5)	0.340 (5)	0.226 (5)
EI-oc-ES	Brainy	0.475 (5)	0.432 (5)	0.544 (6)	0.447 (8)	0.477 (6)	0.340 (6)	0.299 (5)	0.405 (5)	0.302 (8)	0.353 (6)

Table 7: Cohen's kappa for the emotion intensity ordinal classification task

Emotion intensity classification – Pearson (all classes)

Feature	Arabic anger	Arabic sadness	English anger	English fear	English joy	English sadness
ALL*	0.327‡	0.467	0.477	0.470	0.555‡	0.522
-$D^{\dagger}_{250}$		0.467‡	0.490‡	0.467‡		0.497‡
L-en			0.000	-0.090	-0.007	-0.140
L-se		-0.019		-0.023	0.008	-0.030
WN^2_1				-0.055	-0.028	
WE-b				0.001	0.006	
WN^1_1	0.000	0.098				
L-ar		-0.038				
var-CBOW		-0.106				

* Results achieved with all used features for given emotion

† ALL without used LDA feature.

‡ Values used to calculate ablation results.

Table 8: AffectiveTweets feature ablation study

Emotion intensity classification – Pearson (all classes)

Feature	Arabic fear	Arabic joy	Spanish anger	Spanish fear	Spanish joy	Spanish sadness
BoW	-0.013	0.022	0.005	-0.041	0.018	0.003
ChN_1 $t \leq 5$	-0.017	0.024	0.010		0.009	
ChN_2 $t \leq 5$	0.034	-0.037	-0.009		0.018	0.014
ChN_3 $t \leq 5$	-0.053	0.011	0.016	-0.041	0.011	0.005
$ChN_{4,5}$ $t \leq 2$	-0.067	-0.036	-0.008	-0.056	-0.050	-0.011
BoM	-0.022		-0.013		0.017	-0.011
E	0.011		-0.007			
FT	-0.027	-0.008	0.006		-0.004	
BoPOS	-0.015		0.008	-0.010		-0.002
POS-B	-0.008	-0.025	-0.010	-0.013		0.013
BoT	0.017	0.006	-0.003	-0.010		0.018
TF-IDF	-0.017		-0.004		0.009	
NSh	0.010	0.006	-0.011		0.002	-0.008
FW			-0.001		0.002	0.010
LW			-0.007		-0.014	-0.003
TL-B						-0.004
LBoM	0.036		0.000			0.005
V-BoW	-0.006*		-0.005†		0,003‡	

* adverb † adverb, noun, adjective, verb, auxiliary ‡ noun

Table 9: Brainy feature ablation study

Emotion intensity regression – Pearson (all instances)

Feature	English anger	English fear	English joy	English sadness
ALL*	0.640	0.642‡	0.652‡	0.636‡
-$D^{\dagger}_{500}$	0.634‡			
L-en	0.000	-0.044	-0.031	-0.087
L-se		-0.037	-0.010	-0.013
WE-b		-0.020	-0.040	-0.017
	Arabic			
ALL*	0.487	0.559	0.619	0.631
-$D^{\dagger}_{250}$	0.479	0.558	0.604	
L-ar	0.020	0.011	-0.027	-0.027
WN^1_1	0.036			
var-SG	-0.010	-0.244	-0.197	-0.196
	Spanish			
ALL*	0.542	0.688	0.646	0.644
-$D^{\dagger}_{1000}$		0.688		0.639
L-en	0.008	0.006		-0.007
L-es	-0.016	0.005	-0.042	-0.009
L-se		0.002		-0.001
WE-us	-0.021	-0.027	-0.017	-0.030
WN^1_1	-0.033	-0.093		
WN^2_1			-0.050	-0.013
ChN^3_2			-0.006	

* Results achieved with all used features.

† ALL without used LDA feature.

‡ Values used to calculate ablation results

Table 10: AffectiveTweets feature ablation study.

Acknowledgments

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports under the program NPU I and

by university specific research project SGS-2016-018 Data and Software Engineering for Advanced Applications.

References

Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics.

D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Felipe Bravo-Marquez, Eibe Frank, Saif M Mohammad, and Bernhard Pfahringer. 2016. Determining word-emotion associations from tweets by multi-label classification. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, pages 536–539. IEEE.

Fermín L Cruz, José A Troyano, Beatriz Pontes, and F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Tomáš Hercig, Tomáš Brychcín, Lukáš Svoboda, and Michal Konkol. 2016. UWB at SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 342–349. Association for Computational Linguistics.

Michal Konkol. 2014. Brainy: A machine learning library. In Leszek Rutkowski, Marcin Korytkowski, Rafal Scherer, Ryszard Tadeusiewicz, Lotfi Zadeh, and Jacek Zurada, editors, *Artificial Intelligence and Soft Computing*, volume 8468 of *Lecture Notes in Computer Science*, pages 490–499. Springer International Publishing.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, *SEM @ACM 2017, Vancouver, Canada, August 3-4, 2017*, pages 65–77.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Saif Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016a. Sentiment lexicons for arabic social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Saif M. Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016b. How translation alters sentiment. *J. Artif. Intell. Res.(JAIR)*, 55:95–130.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

M Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, and José M Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Will Monroe, Spence Green, and Christopher D Manning. 2014. Word segmentation of informal arabic with domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 206–211.

A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña López. 2012. Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 3–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, Los Angeles, California. Association for Computational Linguistics.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego, California. Association for Computational Linguistics.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 767–777.

Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. 2012. Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican international conference on Artificial intelligence*, pages 1–14. Springer.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R. El-Beltagy. 2017. AraVec: A set of Arabic word embedding models for use in Arabic NLP. *Procedia Computer Science*, 117(Supplement C):256–265.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France. European Language Resources Association (ELRA).

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *JASIST*, 63(1):163–173.

Xabier Saralegi Urizar and Iñaki San Vicente Roncal. 2013. Elhuyar at tass 2013. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS 2013)*, pages 143–150.

David Vilares, Yerai Doval, Miguel A Alonso, and Carlos Gómez-Rodrıguez. 2014. Lys at tass 2014: A prototype for extracting and analysing aspects from spanish tweets. In *Proceedings of the TASS workshop at SEPLN*.

Mohamed A Zahran, Ahmed Magooda, Ashraf Y Mahgoub, Hazem Raafat, Mohsen Rashwan, , and Amir Atyia. 2015. Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–443. Springer.