Cross-lingual Emotion Detection

Sabit Hassan[1], Shaden Shaar[2], Kareem Darwish[3]
[1]University of Pittsburgh, [2]Cornell University, [3]aiXplain Inc
sah259@pitt.edu, ss2753@cornell.edu, kareem.darwish@aixplain.com

Motivation

- Emotion detection can provide us with a window into understanding **human behavior**.
- Constructing annotated datasets for a new language to train machine learning/deep learning models can be **expensive**.
- Cross-lingual and multi-lingual approaches **leverage** data in **another language** to build models for a target language
- Are emotions uniformly expressed across different **languages and cultures**?
- How much **degradation** in quality is expected?

Dataset

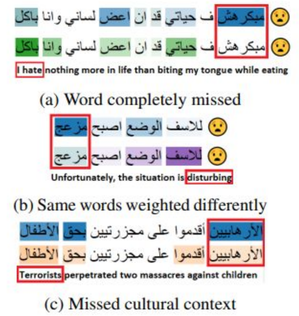
- **SemEval2018** emotion detection dataset (Task 1, E-c). Emotion classes: *joy, sadness, anger, fear, disgust, surprise, trust, anticipation, love, optimism, and pessimism*
- Use **MT** on SemEval data to translate into Arabic and Spanish
- **166K** pairs of English-Arabic parallel tweets (Mubarak et al, 2020)
- **166K** pairs of English-Spanish parallel texts from OpenSubtitles (Tidemann et al., 2012)

Models

- Support Vector Machine (**SVM**):
  - Character n-grams
  - Word embedding representation
  - Universal Sentence Encoder (USE)
  - Character n-grams & embeddings
- Multi-layer perceptron **(MLP)**: USE
- Fine-tuned **BERT**:
  - **Multilingual BERT** (for English)
  - **AraBERT** (for Arabic)
  - **BetoBERT** (for Spanish)
- Parallel corpus approach:
  - Tag English side using **best system** and use translation for training

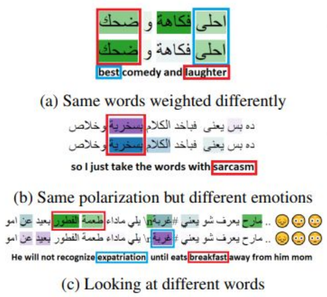
Interpreting Model Differences Using LIME

Mono-lingual model outperforming cross-lingual



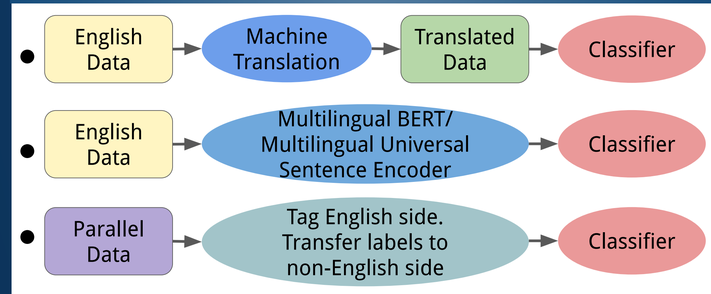
- Cross-lingual model can **miss** an important word
- Both models identified critical words, but **weighted** them *differently*.
- The models can have different **cultural interpretation** for same word

Cross-lingual model outperforming mono-lingual



- Both models identified critical words, but **weighted** them **differently**.
- The models produced tags that were different, but with **matching sentiment**
- The models based their decisions on **different** words

Approach



Contributions

- We show that cross-lingual models (English as the source) can achieve **80–90%** of the relative effectiveness of training SOTA monolingual models for Arabic and Spanish respectively.
- Our multilingual Arabic and Spanish models **notably beat** the best SemEval2018 results.
- We compare models trained on target and source languages using **LIME** to identify potential challenges of cross-lingual approaches.

Experiments and Results

Approach	Model	Features	Cross-lingual J	Cross-lingual F	Cross-lingual A	Combined J	Combined F	Combined A
Best Arabic baseline result			**52.9**	**48.9**	**86.6**			
M	SVM	mUSE	32.4	31.6	79.7	37.9	37.0	83.1
M	MLP	mUSE	35.8	33.4	79.9	41.8	40.5	83.0
M	mBERT	-	20.5	19.5	76.4	46.7	44.6	83.8
T	SVM	C[1-6]	28.0	27.3	81.0	44.4	41.3	85.4
T	SVM	Mzjk	39.3	37.1	82.8	43.7	40.4	85.2
T	SVM	C[1-6]+Mzjk	42.5	40.2	82.5	48.0	45.0	85.6
T	AraBERT	-	**48.1**	**46.3**	**83.8**	**54.1**	**50.8**	**86.2**
P	SVM	C[1-6]	30.4	24.9	74.4	44.1	40.5	85.3
P	SVM	Mzjk	36.5	31.9	74.6	45.1	43.3	84.3
P	SVM	C[1-6]+Mzjk	37.5	32.4	75.2	48.4	46.2	85.3
P	AraBERT	-	36.7	31.6	75.2	53.1	50.5	85.9
M + T	SVM	mUSE	34.3	32.5	80.5	38.4	36.3	83.3
M + T	MLP	mUSE	36.8	34.6	80.7	41.5	38.6	83.0
M + T	mBERT	-	36.0	36.0	79.9	45.9	44.0	83.6
P + M	SVM	mUSE	36.7	33.9	78.9	39.1	37.4	82.8
P + M	MLP	mUSE	37.9	34.9	79.2	41.4	38.6	82.8
P + M	mBERT	-	30.2	26.1	70.6	46.0	44.2	83.3
P + T	SVM	C[1-6]	35.6	34.2	81.2	45.4	41.9	85.1
P + T	SVM	Mzjk	40.5	37.6	81.8	43.0	40.4	84.9
P + T	SVM	C[1-6]+Mzjk	44.0	40.1	82.2	48.0	45.1	85.5
P + T	AraBERT	-	47.7	45.0	82.4	53.1	50.4	85.8
M + P + T	SVM	mUSE	37.2	34.8	80.3	38.1	36.2	82.7
M + P + T	MLP	mUSE	39.3	35.1	79.8	41.5	38.5	82.2
M + P + T	mBERT	-	38.5	36.7	78.8	46.4	44.5	83.6

Approach	Model	Features	Cross-lingual J	Cross-lingual F	Cross-lingual A	Combined J	Combined F	Combined A
Best Spanish baseline result			**52.4**	**53.7**	**89.8**			
M	SVM	mUSE	36.6	36.9	85.7	38.8	37.4	88.4
M	MLP	mUSE	37.4	38.7	84.2	42.5	41.4	88.4
M	mBERT	-	26.5	23.6	83.3	42.3	32.8	88.8
T	SVM	C[1-6]	30.1	28.3	85.3	38.4	37.0	88.3
T	SVM	SBW	23.4	21.9	83.0	22.0	19.6	85.6
T	SVM	C[1-6]+SBW	31.5	30.3	84.1	38.9	37.7	87.9
T	BetoBERT	-	**42.3**	**44.1**	**85.5**	*51.8*	*53.4*	*89.2*
P	SVM	C[1-6]	27.1	31.8	76.6	37.9	38.5	86.2
P	SVM	SBW	24.3	24.8	78.5	23.3	23.6	83.2
P	SVM	C[1-6]+SBW	27.9	32.8	76.5	38.6	39.3	86.1
P	BetoBERT	-	34.1	40.5	77.4	50.5	51.7	88.2
M + T	SVM	mUSE	36.8	35.9	85.3	39.6	38.2	88.2
M + T	MLP	mUSE	37.6	36.4	84.8	42.0	42.0	87.8
M + T	mBERT	-	37.0	33.9	84.6	42.2	34.0	88.4
P + M	SVM	mUSE	36.1	38.1	82.4	36.9	37.4	86.7
P + M	MLP	mUSE	37.0	39.4	82.1	40.1	40.8	87.2
P + M	mBERT	-	32.6	38.6	77.0	47.7	47.6	87.3
P + T	SVM	C[1-6]	33.1	35.6	81.6	40.0	40.0	86.4
P + T	SVM	SBW	24.0	24.3	81.3	23.4	23.3	83.6
P + T	SVM	C[1-6]+SBW	33.6	36.2	81.6	39.7	40.4	86.3
P + T	BetoBERT	-	40.7	44.9	82.3	50.2	52.5	88.0
M + P + T	SVM	mUSE	35.4	37.2	83.8	37.2	37.9	86.7
M + P + T	MLP	mUSE	37.3	39.0	82.9	42.1	41.9	86.7
M + P + T	mBERT	-	38.7	41.3	82.6	47.3	47.3	87.2

Cross-lingual Models: Use English data to train models in Arabic and Spanish

- Use of **Multilingual BERT** and **parallel corpus** do not yield good results
- **Translating** data and fine-tuning BERT in target language is **highly effective**

Combined Models: Use English data + Arabic/Spanish data for training

- **Similar patterns** in performance as cross-lingual models
- **Not large improvement** over training models on **target language alone**

Conclusion

- Cross-lingual models can avail the need to annotate language specific data, and show the **transferability** of emotions across **languages and cultures**
- Using **translated** English training set with **fine-tuned contextual embeddings** led to the best results for both Arabic and Spanish

References

- Mubarak, H., Hassan, S., and Abdelali, A. (2020) .Constructing a bilingual corpus of parallel tweets. In Proceedings of the 13th Workshop on Building and Using Comparable Corpora
- Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)