Enabling Deep Learning of Emotion With First-Person Seed Expressions

Hassan Alhuzali, Muhammad Abdul-Mageed
Natural Language Processing Lab
University of British Columbia
muhammad.mageeed@ubc.ca

Lyle Ungar
Computer and Information Science
University of Pennsylvania
ungar@cis.upenn.edu

Abstract

The computational treatment of emotion in natural language text remains relatively limited, and Arabic is no exception. This is partly due to lack of labeled data. In this work, we describe and manually validate a method for the automatic acquisition of emotion labeled data and introduce a newly developed data set for Modern Standard and Dialectal Arabic emotion detection focused at Robert Plutchik's 8 basic emotion types. Using a hybrid supervision method that exploits first person emotion seeds, we show how we can acquire promising results with a deep gated recurrent neural network. Our best model reaches 70% *F*-score, significantly (i.e., 11%, $p < 0.05$) outperforming a competitive baseline. Applying our method and data on an external dataset of 4 emotions released around the same time we finalized our work, we acquire 7% absolute gain in *F*-score over a linear SVM classifier trained on gold data, thus validating our approach.

1 Introduction

Emotion is a key aspect of human life, and hence emotion detection systems are poised to have a wide array of applications from health and well-being to user profiling, education, and marketing, among others. Compared to prediction of simple valence (i.e., positive vs. negative sentiment) (Wiebe et al., 2004; Pang and Lee, 2004; Balahur and Steinberger, 2009; Liu, 2012; Rosenthal et al., 2017; Yang and Eisenstein, 2017), natural language processing work on emotion recognition still suffers from the bottleneck of labeled data. This is true for the Arabic language. With the exception of Abdul-Mageed et al. (2016) who develop data for Ekman's (Ekman, 1992) 6 basic emotions {*anger, disgust, fear, joy, sadness, surprise*} and another dataset released very recently as part of SemEval 2018 (Mohammad and Kiritchenko, 2018) that focuses on the 4 emotions {*anger, fear, joy, sadness*}, there are no datasets we know of for the language. In this paper, we seek to partially bridge this gap by creating a larger dataset and expanding to Robert Plutchik's list of 8 primary emotions (Plutchik, 1985, 1994) (which adds *anticipation* and *trust* to Ekman's list). In particular, we describe a newly developed, human-labeled dataset using an approach based on emotion phrase seeds from Modern Standard Arabic (MSA) and Dialectal Arabic (DA). In the process, we also seek to enhance the annotation procedure adopted by (Abdul-Mageed et al., 2016) who ask judges to label emotion existence (i.e., whether there is emotion or not) and emotion intensity (i.e., the degree of emotion arousal when an emotion exists) as a single task (rather then two stages). We believe a single stage set up can cause annotator cognitive overload and empirically show how a more simplified, two-stage annotation process yields higher annotator inter-rater reliability.We then proceed to show the utility of exploiting data acquired with our method to develop emotion detection models, including *supervised*, *distant supervised*, and *hybridly-supervised* (i.e., a mixture of supervised and distant supervised). We also validate our method of data acquisition on an external dataset (i.e., (Mohammad and Kiritchenko, 2018)), further proving its usefulness in capturing emotion signal. Finally, training on machine translation (MT) data, we acquire initial results that may be suggesting emotion does not translate (i.e., it may not be possible to successfully build emotion detection systems using MT).

Overall, we offer the following contributions: (1) We extend a first-person seed phrase approach introduced by (Abdul-Mageed et al., 2016) for emotion data collection from 6 to 8 emotion categories, and improve on the annotation procedure, acquiring higher agreement between the judges, (2) we introduce a new dataset for MSA and DA

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 25–35
New Orleans, Louisiana, June 6, 2018.

emotion that is over double the size of their data (i.e., 7,268 vs. 2,984 tweets), (3) we introduce a hybrid supervision method and apply it to develop promising emotion detection models using a powerful deep gated recurrent neural network (GRU), and (4) we explore the utility of MT in the context of emotion detection, hoping our data-driven findings will lead to work enhancing our understanding of emotion.

The remainder of the paper is organized as follows: Section 2 is a review of related work. Section 3 is an overview of the different datasets acquired and used in our work. Section 4 is a description of both the first-person seed phrase approach to data acquisition and the annotation study we performed. Section 5 is about our methods, and 6 is where we introduce our models and describe negative experiments with MT. We conclude in Section 7.

2 Related Work

There is a small, but growing, body of NLP literature on emotion. A number of papers have focused on creating datasets for emotion detection. The SemEval 2007 Affective Text task (Strapparava and Mihalcea, 2007) focused on emotion annotation and classification where a dataset of 1,250 news headlines was human labeled with the 6 basic emotions of Ekman (Ekman, 1972) and provided to participants. Similarly, Aman and Szpakowicz (2007) describe an emotion annotation and classification task on blog post data of 4,090 sentences. The data were collected with identified emotion seeds words. Aman and Szpakowicz (2007) point out that the annotators received no training, but were given samples of annotated sentences to illustrate Ekman's 6 types of emotions. Annotators also labeled the data for *mixed-emotion* and *no-emotion*. In addition, annotators were required to assign emotion intensity tags from the set {*low, medium, high*} to all emotion-carrying sentences (thus excluding sentences tagged with *no-emotion*). Our work differs from these in that we focus on Arabic and the Twitter domain.

A number of works use emotion hashtags (e.g., *#happy, #sad*) as a way of automatically labeling data for emotion (i.e., **distant supervision**) (Mintz et al., 2009). These include Mohammad (2012); Mohammad and Kiritchenko (2015); Wang et al. (2012); Volkova and Bachrach (2016); Abdul-Mageed and Ungar (2017). For example, Mohammad (2012) collects a corpus of 50,000 tweets using seed words corresponding to the 6 Ekman emotions and exploits it for building emotion models. More recently, Mohammad and Bravo-Marquez (2017) label a dataset of 7,097 tweets with emotion intensity tags for the four emotions {*anger, fear, joy, sadness*} using a method they refer to as *best-worst annotation* (Kiritchenko and Mohammad, 2016). They describe the method as producing reliable labels.

In a similar vein, Wang et al. (2012) collect a large emotion corpus (N= 5 million) for 5 of Ekman's 6 basic emotions (skipping *disgust*), but adding *love* and *thankfulness* using a seed set of 131 hashtags representing these emotions. The authors then randomly sample 400 tweets and label them manually with a tag from the set *relevant, irrelevant*. Abdul-Mageed and Ungar (2017) also collect a large dataset of English tweets using 665 hashtags representing 24 different types of emotions. The authors also perform a manual annotation study showing the utility of using hashtags as labels. Other work includes Yan and Turtle (2016) who use crowdsourcing and lab-controlled conditions to label a dataset of 15,553 tweets that they then exploit to build baseline models. Related to our work is also scholarship on **mood** (Nguyen, 2010; De Choudhury et al., 2012) which also depend on collecting data using seed words. Our work also falls under distant supervision, but is different in that we use seed expressions, rather than hashtags. Our data collection method is most similar to Abdul-Mageed et al. (2016), who also use phrase seeds to acquire tweets for Ekman's 6 basic emotions, but we extend the work to 8 emotions, expand the list of seed expressions used, improve on the manual annotation study, and empirically validate the method on the practical emotion modeling task both on our data and on an external dataset. Our work also has affinity to works on Arabic text classification (Abdul-Mageed et al., 2011; Refaee and Rieser, 2014; Abdul-Mageed et al., 2014; Nabil et al., 2015; Salameh et al., 2015; Abdul-Mageed, 2017, 2018; Alshehri et al., 2018; Abdul-Mageed et al., 2018), but we focus on emotion.

3 Data

Building LAMA: We collect a dataset of Arabic tweets from the Twitter public stream ex-

ploiting the Twitter API [1] using a seed set of emotion-carrying expressions following Abdul-Mageed et al. (2016). More specifically, we use a list of seeds for each of the Plutchik 8 primary emotions from the set: {*anger, anticipation, disgust, fear, joy, sadness, surprise, trust*}. As such, we add *anticipation* and *trust* to the 6 categories Abdul-Mageed et al. (2016) work with. In this approach, we collect all tweets where a seed phrase appears in the tweet body text. Note this approach is only conditioned on a given phrase existing in the tweet text as captured by a regular expression. Each phrase is composed of the first person pronoun انا (Eng. "I") + a seed word expressing an emotion, e.g., فرحان (Eng. "happy"). We also follow Abdul-Mageed et al. (2016) in choosing the seed expressions such that they capture data representing Modern Standard Arabic (MSA) as well as Dialectal Arabic (DA). For wider coverage, we expand Abdul-Mageed et al. (2016)'s seeds from 23 to 48 expressions and only include seeds based on complete agreement between two native speakers of the language. From each of the 8 emotion categories, we select $1,000$ tweets with seeds from our list for annotation (total =$8,000$). We ask annotators to manually remove any duplicates in the data, yielding a total of $7,268$ tweets, which we refer to as **LAMA**. To validate this phrase-based approach for emotion data collection, we ask 4 native Arabic speakers to manually label LAMA.

LAMA-DIST: The rest of our dataset acquired with the same seed approach comprises $405,588$ tweets that we automatically clean using a strict pipeline: We remove all re-tweets, use the Python library *pandas* [2] "drop_duplicates" method to compares the tweet texts of all the tweets after normalizing character repetitions [all consecutive characters of >2 to 2] and user mentions (as detected by a string starting with an "@" sign). We then only keep tweets with a minimal length of 5 words. This procedure leaves us with a total of $182,690$ tweets. We call these **LAMA-DIST**.

DINA: We acquire the **DINA** dataset from Abdul-Mageed et al. (2016) and use it in our experiments as we describe in 6.4.

MT-DIST: We use Google Translate to convert the Twitter English dataset from Abdul-Mageed and Ungar (2017) into Arabic and exploit the data to explore the utility of using MT for emotion detection. The data are collected using hashtags representing the same 8 primary emotions we work with. Similar to e.g., Mohammad (2012); Wang et al. (2012) the tweets only involve emotion hashtags occurring in the end of tweets, that have a minimal length 5 words, and tweets with URLs, retweets, etc. are filtered out. An annotation study was performed by the authors (i.e., (Abdul-Mageed and Ungar, 2017)) to validate use of hashtags in this dataset, with 'substantial' inter-annotator agreement over $>5,000$ randomly sampled tweets. In total, we use $756,663$ tweets that we translate into Arabic. We refer to this dataset as **MT-DIST**.

SE-18: We use the SemEval 2018 (Mohammad and Kiritchenko, 2018) Arabic data **(SE-18)** developed for the 4 emotion categories {*anger, fear, joy, sadness*}. Since SE-18 is recently released, only the training and development splits are available. The dataset was collected using emotion-related words and comprises a total of $4,037$ tweets. Table 1 provides statistics of the various datasets we exploit in our experiments. We now turn to describing the annotation study we performed on LAMA to validate our first-person phrase seeds approach.

4 Annotation

4.1 Background

The goal of the annotation is to identify tweets carrying the category of emotion expressed by a given phrase from a set of phrase seeds related to each type of emotion. Conceptualized from this perspective, the annotation process is intrinsically a relevance task where a tweet is judged as *relevant* (i.e., carrying the single emotion expressed by the seed phrase) or *irrelevant* (i.e., carrying no emotion at all or >1 emotion type). Additionally, our goal is to identify the intensity of the emotion in relevant tweets: Given a tweet carrying a single emotion, we direct it to one of three intensity bins. As such, we provide annotators with a tweet where one seed phrase occurs and ask them to approach the task as a two-stage process. In stage one, annotators apply a binary decision using tags from the set {*relevant, irrelevant*}. In stage two, they apply an emotion intensity tag from the set {*low, medium, high*} to all those data points where a *relevant* label was assigned in the first stage. Again, note that we instruct annotators that assigning the

[1] https://dev.twitter.com/.
[2] http://pandas.pydata.org/.

Emotion	DINA		LAMA		LAMA-DINA		LAMA-DIST		MT-DIST		SE-18	
	#	%	#	%	#	%	#	%	#	%	#	%
anger	413	0.14	634	0.09	1,047	0.10	3,650	0.02	45,974	0.06	1,027	0.25
anticipation	–	–	934	0.13	934	0.09	24,673	0.14	24,354	0.03	–	–
disgust	449	0.15	621	0.09	1,070	0.10	2,479	0.01	51,452	0.07	–	–
fear	487	0.16	951	0.13	1,438	0.14	28,332	0.16	65,533	0.09	1,028	0.25
joy	476	0.16	888	0.12	1,364	0.13	55,288	0.3	395,251	0.52	952	0.24
sadness	481	0.16	719	0.10	1,200	0.12	27,609	0.15	130,783	0.17	1,030	0.26
surprise	499	0.17	668	0.09	1,167	0.11	15,108	0.08	34,879	0.05	–	–
trust	–	–	865	0.12	865	0.08	25,550	0.14	8,437	0.01	–	–
no-emotion	179	0.06	988	0.14	1,167	0.11	–	–	–	–	–	–
total/percent	2,984	1.00	7,268	1.00	10,252	1.00	182,689	1.00	756,663	1.00	4,037	1.00

Table 1: Data statistics. **DINA**: Twitter gold-labeled data from Abdul-Mageed et al. (2016). **LAMA**: Our newly-developed dataset. **LAMA-DINA**: A merged set of LAMA and DINA. **LAMA-DIST**: Data we automatically acquire with first-person expressions. **MT-DIST**: Twitter emotion data from Abdul-Mageed and Ungar (2017), translated from English into Arabic. **SE-18** SemEval 2018 Arabic data from Mohammad and Kiritchenko (2018).

Class	Kappa *(K-Bin)*	Kappa *(K-Int)*	% 2-jdgs
anger	0.53	0.66	0.57
anticip	1.00	0.85	0.99
disgust	0.49	0.57	0.97
fear	1.00	0.78	0.97
joy	0.93	0.79	0.92
sadness	0.91	0.90	0.86
surprise	0.77	0.64	0.68
trust	1.00	0.80	0.93
average	0.83	0.75	0.86

Table 2: Annotation agreement. **Kappa (K-Bin)**: binary, emotion vs no-emotion; **Kappa (K-Int)**: intensity-based, fine-grained annotation; **% 2-jdgs**: % of emotion captured per category with double-annotated data.

label *relevant* means the tweet carries the single emotion expressed by the seed phrase. To illustrate an annotation scenario, given a tweet like أنا فرحان جداً لأني زرت أمي بالأمس (En "I'm so happy I visited mom yesterday"), where the seed phrase أنا فرحان (Eng. "I'm happy") indicative of the emotion type "joy" occurs, judges are asked whether the tweet carries the respective single "joy" emotion (i.e., *relevant*) or not (i.e., it either carries no emotion or more than one emotion and hence is *irrelevant*). Judges are then tasked to assign one of the intensity labels to that specific tweet if it is labeled *relevant* in stage one. We note that our emotion intensity procedure is similar to Aman and Szpakowicz (2007). To illustrate the *irrelevant* class, even though a tweet like كلما سأله المذيع عن حالته، أجاب: أنا مبسوط. (Eng. "Whenever the reporter asked him how he is doing, he answered 'I'm glad'.") has the same phrase أنا مبسوط (Eng. "I'm glad") as the previous example, an annotator may decide it does not overall communicate "joy" (i.e., *irrelevant*). Importantly, cast as a two-stage process, our annotation procedure is simpler than Abdul-Mageed et al. (2016)'s single stage set-up where judges are asked to assign one of 4 labels one of which represents *zero* emotion and the rest represent emotion intensity. We believe a two-stage tagging process reduces annotator cognitive overload. As we explain further in Section 4.2, this simplified set-up may be responsible for us acquiring better inter-annotator agreement (Kappa $(K) = 0.75$) than Abdul-Mageed et al. (2016) (Kappa $(K) = 0.51$).

To enable the annotation process and ensure quality, we prepared an annotation guidelines tutorial in the form of a set of presentation slides explaining the overall task, the different emotion categories, the seed expressions chosen to represent each emotion type, and examples of each category. Annotators attended an initial session where the tutorial was shared with them and an expert with native fluency of several Arabic varieties and full knowledge of the task trained them. We had 4 annotators, all of whom are native speakers of Arabic with graduate education. The judges had high proficiency in MSA and reasonable fluency in DA (several dialects). Annotators were advised to consult with one another, consult online sources, and eventually get back to us on cases where a given dialect was not intelligible. Each of the 4 judges labeled data for 2 emotion types. For inter-rater agreement, we chose a sample of 100 labeled tweets from each of the 8 emotions to be double-tagged by the $5th$ judge. We measure inter-annotator agreement using Cohen's (Cohen, 1960) Kappa and also calculate the percentage of

per-class agreement. We now turn to describing findings from the annotation study.

4.2 Annotation Study

Do Seed Expressions Capture Emotion? The main goal of the annotation task is to acquire emotion carrying data that we can exploit in computational models. Hence, the most significant question we had is: "To what extent can first-person seed expressions help capture emotion-carrying data?". Considering the labels assigned by the judges, it turns out that, on average, two judges (middle column in Table 2) agree to assign the *relevant* tag (i.e., one or another of the emotion intensity tags) 86% of the time, whereas one judge (last column in Table 2)) assigns it 89% of the time. Table 2 also shows that our seeds are stronger cues for presence of the respective emotion in some cases more than others. For example, in the case of *anticipation*, judges decided that 99% of the data are *relevant* (i.e., carry the *anticipation* emotion), compared to 57% of the data in the case of *anger*. We now describe hand-labeling the data for emotion intensity.

Can We Consistently Label Intensity? To answer the question as to whether, and if so to what extent, we can label emotion intensity, we asked judges to assign one of three intensity tags from the set $\{$*low, medium, high*$\}$. As Table 2 shows, on average, judges agree on these fine-grained labels with a Cohen's Kappa *(K)* $= 75\%$, thus reflecting 'substantial' agreement (Landis and Koch, 1977). Observably, we acquire higher inter-annotator agreement (Kappa *(K)* $= 75\%$) than (Abdul-Mageed et al., 2016) (Kappa *(K)* $= 51\%$). As we mentioned earlier, this may be a result of our simplified, two-stage annotation set up where judges assign *relevant-irrelevant* tags before they assign intensity labels.[3] We now turn to introducing our methods.

5 Methods

Deep Gated Recurrent Neural Networks: For our core modeling, we use *Gated Recurrent Neural Networks (GRNNs)* (Cho et al., 2014; Chung et al., 2015). Like Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), GRUs constitute a modern variation of *Recurrent Neural Networks (RNNs)* capable of capturing long-term dependencies while side-walking the problems of vanishing/exploding gradients (Bengio et al., 1994; Pascanu et al., 2013). GRUs are simpler than LSTMs, and tend to run faster usually without sacrificing performance, and so we opt for using them. We run an extensive set of experiments, tuning parameters on our dev data. Once we identified the architecture that worked best on most settings, we fix it across all our experiments. Our GRU architecture is as follows: We use a vocabulary size of $50K$ words, a word embedding vector of 300 dimensions learnt directly from the training data, and an input maximum length of 30 words. We use three hidden GRU layers, each with $1,000$ units[4]. For regularization, we use a dropout (Hinton et al., 2012) of 0.5 after the first hidden layer. We use the Adam (Kingma and Ba, 2014) optimizer, setting our learning rate to 0.001. We use a mini-batch (Cotter et al., 2011) size of 128, and run for 4 epochs. For our loss function, we use categorical cross-entropy.

Baseline classifiers: For comparison, we use an SVM classifier with a linear kernel. Since some of our experiments involve larger datasets than what SVMs can handle within memory bounds, we follow Abdul-Mageed and Ungar (2017) in using 4 additional online classifiers: Multinomial Naive Bayes, Passive Aggressive Classifier, Perceptron, and linear SVMs trained with Stochastic Gradient Descent (SVM-SGD). For these, we use the Python scikit-learn package.[5] With the 5 baseline classifiers, for a fair comparison against the deep network models, we experiment with various lexicalized (i.e., based on *N*-grams and lexical resources) features where we identify the best settings for the value of *N* (we experiment with values from the set $\{$*1,2,3*$\}$ and combinations of these) and various vocabulary sizes (we experiment with values between $20K$ and $80K$). Here, we typically tune these hyper-parameters on the dev splits of each of the three datasets LAMA, DINA, and LAMA-DINA independently. We identify unigrams+bigrams (1g+2g) with a vocab_size (V) $= 50K$ as our best settings, and so we fix these across all experiments.

Evaluation: Since we run with several classifiers, we limit reported results to the harmonic mean of precision and recall: *F*-score (macro-average). Unless otherwise indicated, we typically

[3] While we label intensity in our data, we leave detecting intensity to future work.

[4] Models with as less capacity as 500 units performed only slightly worse in most cases.

[5] http://scikit-learn.org.

use the majority class in the training data of each respective set of experiments as our baseline. We now turn to describing our models under various conditions of supervision.

6 Models

6.1 Supervised Models

Data Splits: We first exploit LAMA and DINA, both of which have gold labels, in a supervised fashion. We split each of these data sets into 80% training (**train**), 10% development (**dev**), and 10% test (**test**) and first learn on each of them independently in a standard way where we train on train, tune performance on dev, and blind-test on test. We also merge the corresponding splits from each dataset (e.g., training set from each to acquire a combined train), forming a unified resource (**LAMA-DINA**) that we then exploit under the same supervised conditions. *We consistently remove all our phrase seeds from the data before we perform any of the experiments, even when we run on external data. This is the case for all the experiments we report in the paper.*

Two-Stage Classification: For all supervised experiments, we have a two-stage classification set-up: (1) *binary* where the models attempt to tease apart the *emotion* from the *no-emotion* categories, and (2) the 8-way *emotion* classification. We now present results with our three data settings.

LAMA: Table 3 shows results of our supervised learning settings in *F*-scores. As the Table shows, for the multiclass task on the 8 emotion categories, the best model on LAMA test is acquired with SVMs. SVMs achieve 63% *F*-score, an absolute gain of 48% over the majority class baseline and 5% higher than GRUs (which performs at 58%). For the binary task (i.e., *emotion, no-emotion*, the highest gains of 93% are with the Perceptron classifier (1% over GRUs), again 6% absolute improvement over the baseline.

DINA: As far as we know, we are the first to develop an Arabic emotion system. As such, there is no previous work to compare to. However, as we mention earlier, we acquire the DINA dataset developed by (Abdul-Mageed et al., 2016) and run experiments on it. As Table 3 shows, both SVMs and GRUs perform best on emotion classification on DINA (both at 54%, which is 36% over the baseline). For binary classification, both the Perceptron and GRUs achieve highest, with 98% (i.e., 4% above the baseline).

LAMA-DINA: As explained in Section 5, we merge the corresponding splits from LAMA and DINA to form a single resource (LAMA-DINA). For emotion classification, as in Table 3, GRUs performs best (59% *F*-score, 43% over the baseline) on LAMA-DINA. For the binary task, GRUs also achieves better than other classifiers, with 94% *F*-score (4% above the baseline).

Emotion Lexica: Again, for a fairer comparison with our deep learning models, we experiment with adding lexicon-based features to our online classifiers: Fixing *N*-grams to $1g + 2g$ and $V = 50K$, we use the translated version of the emotion lexicon EmoLex (Mohammad and Turney, 2013) (which has entries for the 8 emotion categories): We add 1 binary feature based on the lexicon to the *emotion vs. no-emotion* stage and 8 binary features (one feature corresponding to each emotion category) to the *emotion* stage. However, we do not find EmoLex features to help, and so we do not use them in further experiments.[6]

Across all the supervised experiments, for both the binary and 8−way emotion classification tasks, our best models are significantly higher than the respective baselines (i.e., at least $< p = 0.05$).

6.2 Distant Supervision with Seeds

We train exclusively on the LAMA-DIST dataset we acquire with seed expressions (as described in Section 3), directly testing performance on LAMA-DINA test set. Across all classifiers, we use the same hyper-parameters described in Section 5). The current and the next sets of experiments (6.3) are focused on **emotion detection** (the 8 types) and are both reported in Table 4. As Table 4 shows, with more training data, GRUs performs better than all other classifiers (53% *F*-score) and is followed by PAC (42%). GRUs' performance is 23% over the baseline, but 6% less than our best result on the same LAMA-DINA test set reported under full supervised (59%, also acquired with GRUs in Section 6.1). This demonstrates the benefit of our phrase-based approach in absence of gold data. We now turn to investigating the utility of employing distant supervision in a scenario where human-labeled data do exist.

[6] We observe a number of issues with the translated version of EmoLex, but leave analysis of these for future work.

TRAIN	Setting	Class	MNB	PAC	PTN	SVM-SGD	SVM	GRU	# test
Lama	**emotion**	base	0.15	0.15	0.15	0.15	0.15	0.15	–
		avg/total	0.49	0.57	0.48	0.54	**0.63**	0.58	632
	binary	base	0.86	0.86	0.86	0.86	0.86	0.86	–
		emotion	0.92	0.92	**0.93**	0.92	0.92	0.92	632
		no-emotion	0.23	0.29	0.21	0.24	0.33	0.28	94
Dina	**emotion**	base	0.18	0.18	0.18	0.18	0.18	0.18	–
		avg/total	0.45	0.40	0.46	0.42	**0.54**	**0.54**	278
	binary	base	0.94	0.94	0.94	0.94	0.94	0.94	–
		emotion	0.89	**0.98**	0.97	0.92	0.97	**0.98**	278
		no-emotion	0.30	0.42	0.41	0.25	0.47	0.46	19
Lama-Dina	**emotion**	base	0.16	0.16	0.16	0.16	0.16	0.16	–
		avg/total	0.43	0.50	0.45	0.47	0.55	**0.59**	910
	binary	base	0.89	0.89	0.89	0.89	0.89	0.89	–
		emotion	0.93	0.93	0.92	0.92	0.93	**0.94**	910
		no-emotion	0.23	0.25	0.38	0.37	0.33	0.30	113

Table 3: **Binary** (i.e., emotion vs. no-emotion) and **emotion** (i.e., a single 8-way classification task) results under supervised conditions. For space, we only report average results across the 8 categories on this set of experiments.

TRAIN	Emotion	MNB	PAC	PTN	SVM-SGD	GRU	#test
Lama-Dist	base	0.30	0.30	0.30	0.30	0.30	—
	avg/total	0.32	**0.42**	0.39	0.41	**0.53**	910
Lama-D2	base-g	0.59	0.59	0.59	0.59	0.59	—
	anger	0.28	0.57	0.53	0.59	0.66	111
	anticip	0.46	0.57	0.54	0.65	0.68	88
	disgust	0.17	0.54	0.50	0.56	0.69	104
	fear	0.60	0.67	0.66	0.69	0.77	145
	joy	0.39	0.48	0.50	0.55	0.68	131
	sadness	0.37	0.49	0.42	0.47	0.63	121
	surprise	0.52	0.59	0.56	0.61	0.72	120
	trust	0.45	0.57	0.60	0.69	0.73	90
	avg/total	0.41	0.56	0.54	**0.60**	**0.70**	910

Table 4: Results of **distant supervision** and **hybrid supervision** on LAMA-DINA test set. **Lama-Dist**: Twitter data we collect with the same phrase-based approach we use in our annotation study. **Lama-D2**: lama-dina+lama-dist. The 59% base-g for the LAMA-D2 setting is what we acquire with gold data (LAMA-DINA training data) with GRUs. We only show average performance with our **Lama-Dist** training data, for space.

6.3 Hybrid Supervision with Seeds

In this iteration of experiments, we merge LAMA-DIST with the training split of LAMA-DINA (80% of the LAMA-DINA data) to form a single training set **LAMA-D2**. As Table 4 shows, with LAMA-D2 as train, GRUs model performance reaches its highest *F*-score of 70%, an absolute gain of 40% over the majority class baseline (base) and 11% absolute gain over the best emotion gold model on the same combined LAMA-Dina test set, a second reasonable baseline (base-g, acquired with GRUs). *This is the best model we report in this paper, and is* ***a statistically significant gain over base and base-g (< p= 0.01 and < p= 0.05, respectively)***. These results further demonstrate the advantage of our first-person phrase seed approach for emotion detection. Based on the current and previous set of experiments, we find that this specific distant supervision approach is valuable when used alone but even more so when used to augment existing gold data.

6.4 Validation on External Data

We further validate our data acquistion approach and models on an externa dataset. For this, we use the SemEval 2018 **(SE-18)** (Mohammad and Kiritchenko, 2018) dataset comprised of 4 emotions (described in Table 1) that was recently released. We only use our best-performing classifier, i.e., GRUs with the same settings as described in Section 5, in this set of experiments. We train GRUs with 5 training data splits and acquire results in *F*-scores as follows: **(a)** SE-18 (36%), **(b)** Lama-Dina (28%), **(c)** Lama-Dist (39%), **(d)** Lama-D2 (41%), and **(e)** Lama-D2+SE-18 (46%). We report only results with conditions **(a)**, **(d)**, and **(e)** in Table 5. As these results show, using only our automatic data (condition **(c)**), we improve 3% over training with SE-18 (condition **(a)**). When

TRAIN	Emotion	GRU	#dev
SE-18	base	0.26	—
	anger	0.00	149
	fear	0.43	145
	joy	0.64	222
	sadness	0.20	139
	avg/total	**0.36**	655
Lama-D2	anger	0.24	149
	fear	0.39	145
	joy	0.58	222
	sadness	0.35	139
	avg/total	**0.41**	655
Lama-D2+SE-18	anger	0.31	149
	fear	0.41	145
	joy	0.64	222
	sadness	0.41	139
	avg/total	**0.46**	655

Table 5: Experiments on **Sem-Eval 2018 (SE-18)** Arabic data on 4 emotion categories.

we add up our distant supervision and gold data (i.e., with Lama-D2), absolute gain goes up to 5%. Augmenting SE-18 with Lama-D2 gives 46% F-score. This is a whole 10% improvement over SE-18 and 20% absolute gain over the 26% majority class baseline. These significant gains on the SE-18 external dataset further demonstrate the utility of our phrase based data acquisition approach, and the advantage of our models.

6.5 Negative Results with MT

In absence of labeled data, MT can be been used for converting labeled data from a source language (often English) into one or more target languages for classification. Although, to the best of our knowledge, there are currently no attempts to exploit MT for emotion detection, there have been successful efforts on the (conceptually relevant) task of sentiment analysis. Examples of sentiment systems employing MT include Hiroshi et al. (2004) (Japanese), Wan (2008) (Chinese), Brooke et al. (2009); Smith et al. (2016) (Spanish), Mihalcea et al. (2007) (Romanian), and Mohammad et al. (2016) (Arabic). Clearly, MT has its limitations. Hence, whether MT will be as useful for emotion as it proved to be for sentiment is in our view an interesting question. As a first attempt to explore answers, we experiment with the MT-DIST data described in Section 3 under two settings: **(a)** We train exclusively on MT-DIST and test on LAMA-DINA, and **(b)** We merge MT-DIST with the training split of LAMA-DINA to form a single training set that we refer to as **MT-D2**. Again, we use the same settings as described in 5 with both the online classifiers and GRUs, and directly test on LAMA-DINA test set. For space limitations, we do not report the full results from this cycle of experiments with MT here. We do note, however, that we acquire no gains on either of the two settings: With MT-DIST functioning as our training data, the best model we acquire is with GRUs (only at 10% F-score, i.e., 5% less than the baseline). Similarly, with MT-D2 as train, GRUs acquires a best result of 20%, a performance 39% less than the 59% we acquire with GRUs using the LAMA-DINA gold data (reported in Table 3). This shows that MT data hurts emotion classification when used for training.

A full understanding of why it is that MT does not help emotion classification is beyond our current work. However, we hypothesize a number of reasons could account for our findings. Intuitively, MT is in general prone to errors and these could be naturally propagating to our models. In addition, the original Twitter dataset which we convert into MT-DIST is acquired via distant supervision, a regime that may have its own biases and noise. From a theoretical perspective, although early psychological research claimed the universality (i.e., cross-cultural nature) of basic emotions, such work is based on facial expression premises, not language, and are not uncontroversial (Barrett, 2017; Mesquita et al., 2017). We suspect there are cross-cultural variations, even in these primary emotions, that current MT technologies cannot capture. Finally, the fact that our test data involves Dialectal Arabic (a range of varieties Google's production MT models do not currently handle) is in all likelihood responsible for a share of the errors.

7 Conclusion

In this paper, we evaluated the feasibility of automatic acquisition of emotion data from the Twitter domain using an approach based on first-person expressions. We validated the method via a careful, manual annotation study. We then developed successful supervised, distant supervised, and hybrid supervised models exploiting the data and validated our methods on an external dataset. We also explored the utility of using MT for emotion detection, providing initial insights that we hope will ultimately lead to enhanced, cross-cultural understandings of emotion. In the future, we plan to extend our models to different emotion categories and possibly other languages.

8 Acknowledgement

This research was enabled in part by support provided by WestGrid (https://www.westgrid.ca/) and Compute Canada (www.computecanada.ca).

References

Muhammad Abdul-Mageed. 2017. Modeling arabic subjectivity and sentiment in lexical space. *Information Processing & Management* .

Muhammad Abdul-Mageed. 2018. Learning subjective language: Feature engineerd vs. deep models. In *The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT 2018), LREC*.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *LREC*.

Muhammad Abdul-Mageed, Hassan AlHuzli, and Mona Diab DuaaAbu Elhija. 2016. Dina: A multi-dialect dataset for arabic emotion analysis. In *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*. page 29.

Muhammad Abdul-Mageed, Mona Diab, and Mohamed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 587–591. http://www.aclweb.org/anthology/P11-2103.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language* 28(1):20–37.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 718–728.

Ali Alshehri, AlMoetazbillah Nagoudi, Alhuzali Hassan, and Muhammad Abdul-Mageed. 2018. Think before your click: Data and models for adult content in arabic twitter. In *The 2nd Text Analytics for Cybersecurity and Online Safety (TA-COS-2018), LREC*.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*. Springer, pages 196–205.

A. Balahur and R. Steinberger. 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceeding of WOMSA* .

Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5(2):157–166.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*. pages 50–54.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Junyoung Chung, Caglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *ICML*. pages 2067–2075.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46.

Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. 2011. Better mini-batch algorithms via accelerated gradient methods. In *Advances in neural information processing systems*. pages 1647–1655.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! exploring human emotional states in social media.

P. Ekman. 1972. Universal and cultural differences in facial expression of emotion. *Nebraska Symposium on Motivation* pages 207–283.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6(3-4):169–200.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* .

Kanayama Hiroshi, Nasukawa Tetsuya, and Watanabe Hideo. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 494.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Svetlana Kiritchenko and Saif M Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *HLT-NAACL*. pages 811–817.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

Batja Mesquita, Michael Boiger, and Jozefien De Leersnyder. 2017. Doing emotions: The role of culture in everyday emotions. *European Review of Social Psychology* 28(1):95–133.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Annual meeting-association for computational linguistics*. volume 45, page 976.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, pages 1003–1011.

S. Bravo-Marquez F. Salameh M. Mohammad and S. Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*. Association for Computational Linguistics.

Saif M Mohammad. 2012. #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 246–255.

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

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* 31(2):301–326.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. 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.

Mahmoud Nabil, Mohamed Aly, and Amir F Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 2515–2519.

Thin Nguyen. 2010. Mood patterns and affective lexicon access in weblogs. In *Proceedings of the ACL 2010 Student Research Workshop*. Association for Computational Linguistics, pages 43–48.

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

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)* 28:1310–1318.

Robert Plutchik. 1985. On emotion: The chicken-and-egg problem revisited. *Motivation and Emotion* 9(2):197–200.

Robert Plutchik. 1994. *The psychology and biology of emotion.*. HarperCollins College Publishers.

Eshrag Refaee and Verena Rieser. 2014. An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*. pages 2268–2273.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. pages 502–518.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *HLT-NAACL*. pages 767–777.

Laura Smith, Salvatore Giorgi, Rishi Solanki, Johannes C. Eichstaedt, Hansen Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle H. Ungar. 2016. Does 'well-being' translate on twitter? In *EMNLP*.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pages 70–74.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *EMNLP*. Association for Computational Linguistics, pages 553–561.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, pages 587–592.

J. M. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational Linguistics* 30:227–308.

Jasy Liew Suet Yan and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of NAACL-HLT*. pages 73–80.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *TACL* 5:295–307.