Exploring Fine-Tuned Embeddings that Model Intensifiers for Emotion Analysis

Laura Bostan and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{laura.bostan,roman.klinger}@ims.uni-stuttgart.de

Abstract

Adjective phrases like "a little bit surprised", "completely shocked", or "not stunned at all" are not handled properly by currently published state-of-the-art emotion classification and intensity prediction systems which use predominantly non-contextualized word embeddings as input. Based on this finding, we analyze differences between embeddings used by these systems in regard to their capability of handling such cases. Furthermore, we argue that intensifiers in context of emotion words need special treatment, as is established for sentiment polarity classification, but not for more fine-grained emotion prediction. To resolve this issue, we analyze different aspects of a post-processing pipeline which enriches the word representations of such phrases. This includes expansion of semantic spaces at the phrase level and sub-word level followed by retrofitting to emotion lexica. We evaluate the impact of these steps with À La Carte and Bag-of-Substrings extensions based on pretrained GloVe, Word2vec, and fastText embeddings against a crowd-sourced corpus of intensity annotations for tweets containing our focus phrases. We show that the fastText-based models do not gain from handling these specific phrases under inspection. For Word2vec embeddings, we show that our post-processing pipeline improves the results by up to 8% on a novel dataset densely populated with intensifiers.

1 Introduction

Emotion detection in text includes tasks of mapping words, sentences, and documents to a discrete set of emotions following a psychological model such as those proposed by Ekman (1992) and Plutchik (1980), or to intensity scores or continuous values of *valence–arousal–dominance* (Posner et al., 2005). The shared task on intensity prediction for discrete classes proposed to combine both (Mohammad et al., 2018; Mohammad and Bravo-Marquez, 2017a). In this task a tweet and an emotion are given and the goal is to determine an intensity score between 0 and 1.

Especially, but not only in social media, users use degree adverbs (also called intensifiers Quirk, 1985), for instance in "I am *kinda happy*" *vs.* "I am *very happy*." to express different levels of emotion intensity. This is a relevant task: 10% of tweets containing an emotion word are modified with such an adverb in the corpus we describe in Section 3.1. In this paper, we challenge the assumption that models developed for intensity prediction perform well on tweets containing such phrases and analyze which of the established embedding methods Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText embeddings (Bojanowski et al., 2017) performs well when predicting intensities for tweets containing such phrases. We will see that the performance of the popular and fast-to-train Word2vec method can be increased with a simple postprocessing pipeline which we present in this paper.

As a motivating example, the DeepMoji model (Felbo et al., 2017) predicts *anger* for both the example sentences "I am not angry." and "I am angry."[1]. Using the model by Wu et al. (2018) (one of the state-of-the-art intensity prediction models from Mohammad et al. (2018), building their model on top of Word2vec embeddings) we also obtain *anger* as having the highest intensity for both examples. We argue that the models should be more sensitive to the difference between *negations*, *downtoners* and *amplifiers*.

With this paper, we contribute to alleviate this situation in three aspects. Firstly, we provide an analysis of the distribution of degree adverbs (in-

[1] https://deepmoji.mit.edu

Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 25–34
Minneapolis, June 6, 2019.

cluding negations) with emotion words and show that not all such combinations are equally common. Secondly, we perform a crowdsourcing experiment in which we collect scores for different combinations of degree adverbs and emotion adjectives. We use these data, which we make publicly available, as an additional challenging test set for the task of intensity prediction for English. Thirdly, we use a state-of-the-art intensity prediction model (Wu et al., 2018) on this test set and evaluate two methods to improve these predictions, namely the inclusion (Zhao et al., 2018) and n-gram embeddings via À La Carte of additional subword information with Bag-of-Substrings (Khodak et al., 2018). We evaluate based on Word2vec, GloVe and fastText embeddings and show that particularly the first two benefit from these changes, but to different extents.

2 Related Work

2.1 Degree Adverbs in Linguistics

Adverbs that express intensity are named *degree adverbs*, *degree modifiers* or *intensifiers*.[2] The entities they intensify are located on an abstract scale of intensity (Quirk, 1985). The intensifiers that scale upward are named amplifiers and are further categorised as maximizers, such as "completely" and "totally" or boosters, such as "really" or "truly". Those that scale downward are called downtoners and are further classified as approximators, such as "almost" or "kind of", compromisers, such as "fairly", "pretty" and "quite", diminishers, such as "slightly" and "a bit", and minimizers (Quirk, 1985; Paradis, 1997; Nevalainen and Rissanen, 2002, *i. a.*). Further distinction of degree modifiers is concerned with the fact that there are intensifiers that imply boundaries, such as "totally", "fully", and "completely" and those that do not, such as "very", "utterly", "pretty" (Paradis, 1997, 2001, 2000). Finally, in the context of discourse, there is the property of expressing focus, which is present in the so-called *focus modifiers*, such as "only" and "just", which are also further classified in additives, such as "also" and "too" and restrictives, such as "only" and "merely" (Quirk, 1985; Athanasiadou, 2007).

English degree modifiers have also long history of research in English studies and more generally in Language Studies. Most English studies focus on the incidence and distribution of these adverbs in different corpora, e.g. Peters (1994) study letters from Early Modern English and shows the how the distributions of boosters change across time. Nevalainen (2008) study the social variation in intensifier use, with a focus on the suffix *-ly*. More recently, Napoli and Ravetto (2017) collect a volume of papers that explore the process of intensification following a corpus-based, cross-linguistic and contrastive approach. The volume contains various works on the variation in the distribution and incidence of the intensifiers based on sociolinguistic features and in a diachronic fashion. The work brings in attention intensification in ancient languages as well as modern languages.

A more recent work investigates the differences in the use of intensifiers and considers English speech of adults and teenagers as corpus. It explores two maximizers in-depth, namely "absolutely" and "totally" and shows that those prove to be more "flexible" in the language used by teenagers (Pertejo and Martínez, 2014).

2.2 Modifiers in the context of Sentiment and Emotion Analysis

In the context of sentiment analysis the discussion of intensifiers and negations has gained quite some attention, since those are primarily markers of subjectivity (Athanasiadou, 2007).

Negations, and in particular negation cue detection (with the goal of scope recognition) have been the research interest of Councill et al. (2010) and Reitan et al. (2015), who use a lexicon for negation cue detection and a linear-chain conditional random field for scope recognition. In the area of distributional semantics, the investigation of word vectors with a focus on negated adjectives (Aina et al., 2018) is complementary to our work with regards to negation in terms of the methods and data used. Following this approach, one could build a distributional semantic model whose vocabulary includes the modified phrases. In practice, each occurrence of a modified adjective by a degree adverb could be treated as a single token (*e. g.* "not happy" would be represented as "not_happy"). For a general overview of modality and negation in computational linguistics we refer the interested reader to the work by Morante and Sporleder (2012).

Furthermore, Zhu et al. (2014) study the effect of negation words on sentiment and evaluate a neural composition model. Kiritchenko and Mohammad (2016a) create a sentiment lexicon of phrases that include modifiers such as negators, modals, and

[2] In this paper, we will use these terms interchangeably.

degree adverbs. The phrases and their constituent words are annotated manually with the same annotation procedure we will discuss in detail. We follow this work closely and apply the same procedures in the context of emotion analysis.

Dragut and Fellbaum (2014) study the effect of intensifiers on the sentiment ratings and shows that the degree adverbs do not carry an inherent sentiment polarity but alter the degree of the polarity of the constituents they modify.

We argue that there is not enough work on transferring the methods used in sentiment analysis to the more fine-grained analysis of emotions, except for Strohm and Klinger (2018), who limit themselves to analysis and do not apply state-of-the-art prediction models for handling degree adverbs, and Carrillo-de Albornoz and Plaza (2013) who consider modified emotions but predict sentiment.

3 Methods

In the following, we explain how we create the data sets for our analysis (Section 3.1) and then how we set up the experiments to measure the impact of À La Carte and Bag-of-Substrings on the modified phrases (Section 3.2).

3.1 Data Collection and Annotation

As a basis of our work, we create a compositional emotion lexicon for English Twitter and retrieve crowdsourced ratings using *Best-Worst Scaling* (Louviere et al., 2015; Kiritchenko and Mohammad, 2016b). We show later that these ratings are by and large independent of context and can therefore be interpreted as a labeled emotion lexicon of compositional phrases.[3]

3.1.1 Data Collection

Each query we use to retrieve tweets consists of a pair of an adjective with one or a combination of several degree adverbs (intensifiers (including amplifiers and downtoners) and negations), for instance "not at all surprised" or "not very happy". We first generate a comprehensive list by mapping each of Ekman's fundamental emotions (Ekman, 1992) to their corresponding adjective *sad, happy, disgusted, afraid, surprised, angry* and augment this list to 333 emotion adjectives and their synonyms from the Oxford Dictionary of English (Ehrlich, 1980), the New Oxford American Dictionary (Stevenson and Lindberg, 2010) and Macmillan Online English Dictionary[4] and further filter this list to 43 entries which are intersubjectively agreeable. This filter step is performed via crowdsourcing on Prolific[5], in which we asked native speakers of English which emotion is the closest to each synonym. We only keep those synonyms where all annotators agreed. The inter-annotator agreement is $\kappa = 0.63$ (Fleiss' κ over 9 annotators).

The list of degree modifiers is a combination of Quirk (1985); Paradis (1997); Strohm and Klinger (2018). From the cartesian product of degree modifiers with emotion adjectives, we keep those which we find at least 10 times in the general Twitter corpus we discuss below. That leads to 266 phrases.

We base our analysis on a set of 32 million tweets obtained from Twitter with the official API between March 2006 and October 2018, using a combination of diverse search terms corresponding to isolated emotion word synonyms, those in combination with degree adverbs, and frequent hashtags. We filter out retweets and full quotes, tweets with more than 140 characters and those with less than 10 tokens, as well as those consisting of more than 30% hashtags, links, or usernames, which we replace by generic respective tokens otherwise. Tweets with more than 30% of non-ASCII characters are also removed.

3.1.2 Annotation Procedure

For each tweet (t) and emotion (e) we obtain emotion intensity scores $s_{t,e} \in [\ \ 1, 1]$ via *Best-Worst Scaling* (BWS, Louviere et al., 2015). In general with BWS, the annotators are shown a subset of a number of items from a list and are asked to select the *best and worst* items (or most and least some given property of interest). Within our study, we show four items at once to the annotators. In a first setting, we show them four tweets that contain the queries we want to have scores assigned for. In a second setting, we show them only the queries without the context (the tweet) in which they were found. In both scenarios, the annotators need to select the tweet or the query with the highest and lowest intensity of each emotion.

These groups of tweets are sampled under following constraints that have been empirically

[3] Our data is available at https://www.ims.uni-stuttgart.de/data/modifieremotion.

[4] http://www.macmillandictionaries.com/dictionary-online/

[5] https://prolific.ac

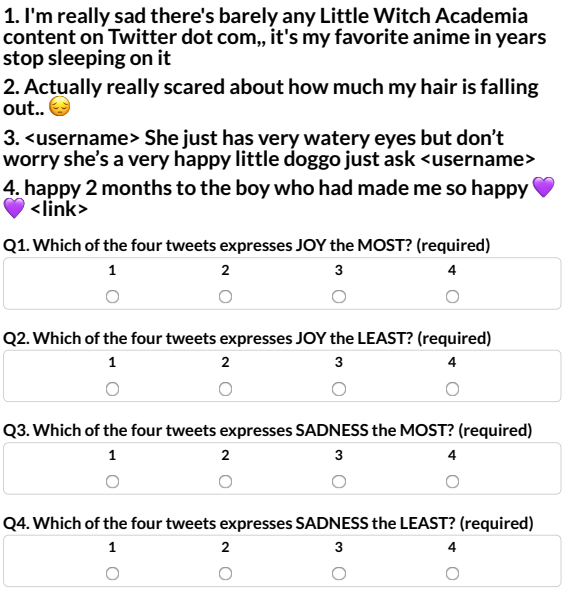

Figure 1: An example of what contributors see on the Figure Eight Platform. The 4 sentences shown are an example of a group of four tweets the contributors have to annotate. The questions Q1 to Q4 that follow below are a subset of the questionnaire.

proven to lead to reliable scores (Kiritchenko and Mohammad, 2016c), resulting in 532 samples (twice the amount of queries): (1) no two samples have the same four queries (in any order), (2) no two queries within a sample are identical, (3) each query occurs in 8 (± 1) different samples, (4) each pair of queries appears in the same number of samples. We perform two annotation experiments on the crowdsourcing platform Figure Eight[6]: In Experiment 1, we present the whole tweet to the annotator, in Experiment 2, we only show the query phrase. This enables us to evaluate the importance of context, shown in Section 4.2. Each sample was annotated by three contributors that confirmed to be English native speakers.

3.2 Adaptations of Embeddings

In the following, we discuss the three methods to improve the embeddings and later to test if these improvements add additional information with respect to intensifiers for emotion analysis. The evaluation will be on the downstream task of emotion intensity prediction.

We focus on subword-level information and phrase-level information, as those, presumably, capture intensity information.

[6] https://www.figure-eight.com

3.2.1 À La Carte

With this method we learn a representation of yet unseen phrases within an embedding space through a linear transformation of the average of the word embeddings in the feature's contexts. The method constructs a representation for a new phrase given a set of contexts where this phrase occurs in.

Given our Twitter corpus $\mathcal{C}_w$ consisting of contexts of words w and the pre-trained word embeddings $\mathbf{v}_w \in \mathbb{R}^d$, of dimension d, our goal is to construct a representation $\mathbf{v}_q \in \mathbb{R}^d$ of a query q given a set $\mathcal{C}_q$ of contexts it occurs in.

We learn the transform $\mathbf{A} \in \mathbb{R}^{\mathbf{d}\times\mathbf{d}}$ that can recover *existing* word vectors $\mathbf{v}_w$ via *linear regression* by summing their context embeddings

$$\mathbf{v}_w \approx \mathbf{A}\left(\frac{1}{|\mathcal{C}_w|}\sum_{c\in\mathcal{C}_w}\sum_{w'\in c}\mathbf{v}'_w\right). \quad (1)$$

Using the learned transformation matrix $\mathbf{A}$ we can embed any new query $\mathbf{v}_q$ in the same semantic space as the pre-trained word embeddings via

$$\mathbf{v}_q = \mathbf{A}\left(\frac{1}{|\mathcal{C}_q|}\sum_{c\in\mathcal{C}_q}\sum_{w\in c}\mathbf{v}_w\right). \quad (2)$$

3.2.2 Bag-of-Substrings

BoS generalizes pre-trained semantic spaces to unseen words. The established approach to represent word phrases or sentences is to take a bag of words of word embeddings.

BoS achieves its goal by first learning a mapping between the subwords present in each word and its corresponding pre-trained vector. Then, by using this learned subword transformation, the model is able to generate new representations for any new word as a set of its character n-grams. For us this is relevant, since we can consider our focus phrases to be character n-grams instead of word n-grams.

Formally, the representation for a word $\mathbf{v}_w$ from the lookup table V (which stores the embeddings of dimension d for each possible substring of length within a range) is:

$$\mathbf{v}_w = \frac{1}{|\mathcal{S}_w|}\sum_{t\in\mathcal{S}_w}\mathbf{v}_t, \quad (3)$$

where $\mathcal{S}_w$ is the set of each possible character n-grams of length within a given range over w and $\mathbf{v}_t$ is the vector in V indexed by t.

The model views the vector of a phrase as the average vector of all its substrings, which are trained

by minimizing the overall mean squared loss between the generated and given vectors for each word:

$$\min_{V} \frac{1}{|W|} \sum_{w \in W} l\left(\frac{1}{|\mathcal{S}_w|} \sum_{t \in \mathcal{S}_w} \mathbf{v}_t, \mathbf{u}_w\right) \quad (4)$$

where $\mathbf{u}_w \in \mathbb{R}^{d \times |W|}$ are the target vectors of the dimension d over the vocabulary W and $l(\mathbf{v}, \mathbf{u}) = \frac{\mathbf{1}}{\mathbf{2}} \|\mathbf{v} \quad \mathbf{u}\|_{\mathbf{2}}^{\mathbf{2}}$.

After training, similarly to the previous method, one can use the learned space to generate a new word vector $\mathbf{v}_q$ as the average of the vectors of all of its substrings through Equation 3.

Since BoS produces vectors for unknown words from vectors of substrings of characters contained in it, this allows to build vectors for misspelled words and concatenation of words. Particularly on Twitter data, we benefit from getting a representation for phrases like "sooooexcited:)", "verrry cheerful", "soo unhappy:(". Relevant for our analysis is that BoS uses special characters to mark the start and the end of the word and thus helps the model to distinguish morphemes that occur at different word parts, like prefixes or suffixes. Through that we learn to distinguish morphemes like "un-", "-er" and "-est" that are part of our focus phrases.

Note that this method uses the same idea as in fastText (Bojanowski et al., 2017), but is for our case computationally more efficient, since the BoS model is trained directly on top of pre-trained vectors, instead of predicting over text corpora.

3.2.3 Retrofitting

We use the method of retrofitting existing embeddings (Faruqui et al., 2015) in order to enrich word vectors using synonymity constraints provided by semantic lexicons. The algorithm learns the word embedding matrix $A = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ with the objective function:

$$(A) = \sum_{i \in V} [\alpha_i ||\mathbf{v}_i \quad \hat{\mathbf{v}}_i||^2 + \sum_{(i,j) \in E} \quad _{ij} ||\hat{\mathbf{v}}_i \quad \hat{\mathbf{v}}_j||^2] \quad (5)$$

where an original word vector is $\mathbf{v}_i$, its synonym vector is $\mathbf{v}_j$, and inferred word vector is $\hat{\mathbf{v}}_i$.

Our lexicon of synonymity constraints was automatically constructed from the data we collected in Section 3.1.1 by adding an entry for each emotion adjective with its synonyms crowdsourced as previously described. We also added entries for the phrases in the lexicon for retrofitting, as follows, for each emotion phrase according to the intensifiers classification described in Section 2.1. For instance, "not happy" had as an entry in the lexicon the phrases "unhappy" and "not happy at all" while "completely cheerful" had in its entry phrases like "totally cheerful", "totally happy", "completely happy", among others (since "completely" and "totally" are in the same class). We apply retrofitting on phrases which origin from the extension of the space with À La Carte.

Focus phrase	joy	sadness	anger	fear	surprise	disgust
so happy	+.73	−.43	−.50	−.51	−.10	−.66
not happy	−.52	+.41	+.02	−.16	−.11	+.17
kinda happy	+.53	−.70	−.67	−.55	−.47	−.76
so sad	−.50	+.66	+.04	+.13	−.16	+.03
not sad	+.55	−.60	−.57	−.55	−.45	−.52
kinda sad	−.41	+.62	−.02	+.02	−.18	+.02
so angry	−.39	+.26	+.86	+.21	+.02	+.63
not angry	+.40	−.36	−.82	−.17	−.27	−.45
kinda angry	−.80	+.68	+.84	+.32	+.08	+.64
so scared	−.07	+.15	−.21	+.83	+.15	−.13
not scared	+.35	−.35	−.28	−.66	−.53	−.33
kinda scared	+.03	+.10	−.03	+.71	−.13	−.13
so surprised	+.34	−.27	−.09	+.02	+.81	.00
not surprised	+.60	−.56	−.50	−.60	−.83	−.60
kinda surprised	+.37	−.37	−.17	+.01	+.72	−.20
so disgusted	−.11	−.02	+.30	+.16	+.33	+.88
not disgusted	+.42	−.39	−.42	−.36	−.36	−.84
kinda disgusted	−.16	+.08	+.41	−.01	+.08	+.80

Table 1: Example queries with their BWS crowdsourced scores for the modifiers "so", "kinda" and the negation "not". For every focus phrase we have an intensity score between 1 and +1 for each emotion. The focus phrases are shown in groups made around the emotion adjectives.

4 Results

In the following, we explore the Twitter corpus described previously, the results of the BWS annotation of the pairs of degree adverbs and adjectives, and finally we discuss our experimental setting and evaluation on the downstream task of emotion intensity prediction on the different embedding adaptation methods.

4.1 Corpus Analysis

The Twitter corpus described in Section 3.1 contains 34,297,941 tweets out of which 2,948,397 contain emotion phrases. Most dominant are am-

plifiers (49%) followed by downtoners (24%) and negations (19%). Only 8% contain the emotion adjectives in superlative or comparative.

Figure 2 shows how often the top 30 modifiers are used with adjectives from the basic set of emotions. We see that *disgust* is rarely downtoned and *anger*, *sadness*, and *surprise* are amplified most often. *Joy* and *fear* are relatively equally amplified, with *joy* being more negated and *fear* being more downtoned. The amplifiers "so" and "really", as well as the downtoners "just" and "kind of/kinda" are frequently used. The downtoner "just" is the most frequently used downtoner and acts at times as an amplifier, which could explain its frequent use. We hypothesize that this is due to their use as fillers and their grammaticalization (cf. Tagliamonte, 2006). Most frequently downtoned emotion is *surprise* (which is often used in phrases like "a little surprised", "quite surprised", "a bit surprised").

In Figure 3 and Figure 4 we observe that the use of modifiers with respect to an emotion vary a lot within the same class of modifiers among both more frequent and less frequent modifiers. In Figure 3, we observe that the focus modifier "only" scales downward *surprise* the least, while all the other "true" scaling adverbs are more impactful. *Sadness* is the emotion that is mostly expressed through the focus adverb "only" in this setting. The figure also (implicitly) shows that certain modifiers prefer certain adjectives, e.g. the adjectives that express *disgust*, such as "disgusted" is mostly modified by "absolutely", "truly", "utterly", "pretty" and not by "extremely", "incredibly" or "only". This distinction show the "harmony" between adjectives and degree adverbs (Quirk, 1985).

Looking in more depth into the most frequent used amplifiers and downtoners in Figure 4 we see that among the top used amplifiers "so", "really", "very" we find that *joy*, *anger*, and *disgust* prefer "so" over "really" and "very", the emotions *fear* and *surprise* prefer "very" over "so" and "really" and *sadness* is modified rather equally by the three amplifiers. Between the downtoners "kind of" and "kinda" there is a notable difference in use for *sadness*, *fear* and *anger*, with "kind of" being prefered over "kinda" in the context of *sadness*, with the opposite holding true for *fear*.

4.2 Annotation Analysis

Table 1 shows examples of phrases annotated with real-valued scores following the annotation procedure described in Section 3.1. We see that we have scores for each phrase in the context of each emotion. For instance, "kinda surprised" has the score .37 for *sadness* and +.17. We observe that the negation "not" paired with any emotion adjective, excluding *happy* obtains a positive score for *joy*, and a negative score for every other emotion. The phrase "not happy" obtains a negative score of only .52. In the complete annotation results we include as negations also the phrase "not happy at all", which in this case gets closer to the lower limit of the potential scores.

	Spearman's rank correlation		
Emotion	*w/ context*	*w/o context*	*between*
anger	.84	.82	.88
fear	.84	.73	.81
joy	.90	.86	.91
sadness	.90	.86	.88
surprise	.71	.71	.81
disgust	.86	.86	.88
average	.84	.80	.86

Table 2: Split-half reliabilities and Spearman's rank correlation between these settings.

We measure the reliability by randomly dividing the sets of 4 responses to each question into two halves and comparing the Spearman rank correlation coefficient between the two sets (Kiritchenko and Mohammad, 2016b). Both with and without having access to context, the annotators mostly agree regarding their annotations, as Table 2 shows in the first two columns. Lowest reliability is achieved for *surprise*, with .71 Spearman's rank correlation and the highest for *joy* and *sadness* (.9). The reliability drops most when context is not available for fear (by 11 percentage points).

Figure 5 shows the distribution of the scores assigned through the annotation per emotion. We observe that *disgust* is mostly amplified and rarely negated (only once). The outliers in each boxplot mostly correspond to negated phrases.

4.3 Embedding Adaptations

Figure 6 summarizes our experimental setup. We build on top of pretrained embeddings obtained with Word2vec (Mikolov et al., 2013) (300d, negative sampling, Google News corpus), fastText (Bojanowski et al., 2017) (300d, news corpora), or GloVe (Pennington et al., 2014) (300d, Com-

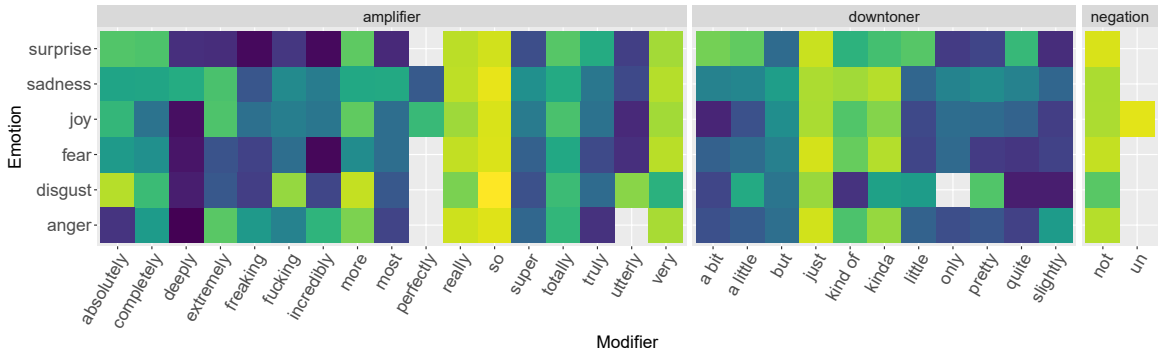

Figure 2: Relative frequencies of the most common 30 modifiers in the Twitter Corpus (from dark (infrequent) to yellow (frequent).

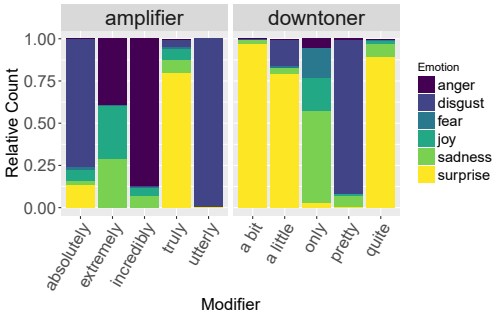

Figure 3: Amplifiers and downtoners that vary the most in use with regards to emotion

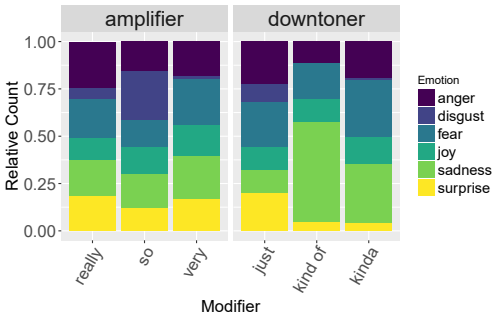

Figure 4: Most frequent three amplifiers and downtoners used across all emotions and their variation with respect to emotion.

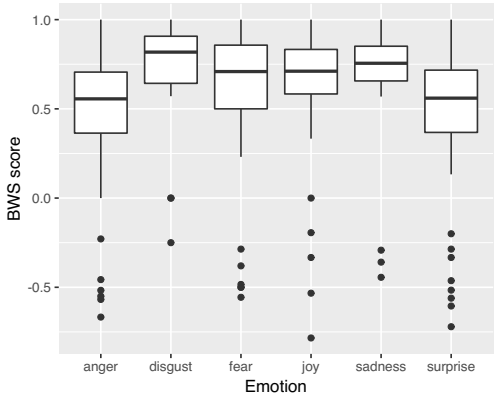

Figure 5: Distribution of the aggregated emotion scores obtained by applying the counting procedure BWS

mon Crawl). Each embedding is then optionally augmented with phrase and subword embeddings and fed into a CNN-LSTM model as proposed by Wu et al. (2018), trained on the Affect in Tweets Dataset used at Sem Eval 2018 Task 1 (Mohammad et al., 2018). Their system achieved an average Pearson correlation score of 0.722, and ranked 12/48 in the emotion intensity regression task.

Table 3 shows Spearman's rank correlation between the predicted intensity scores and the emotion scores obtained in the annotation of our Twitter corpus or the EmoInt data (Mohammad and Bravo-Marquez, 2017b).

The fastText-based models underperform constantly on our Twitter dataset. For GloVe embeddings, À La Carte (ALC) and Bag-of-Substrings (BoS) lead to a substantial improvement, of 7pp (see Table 3, G vs. G+ALC) and 8pp (G vs. G+BoS) over the baseline of using the pretrained embeddings unchanged. On Word2vec embeddings BoS and ALC show the same improvement of 7pp (W2V vs. W2V+ALC/BoS).

While on average, ALC and BoS can only substantially contribute based on GloVe and Word2vec, this is not the case for individual emotions. For Word2vec, sadness figures to be particularly challenging, leading to an overall comparably low performance. Most importantly, we observe that our extensions of the semantic spaces do not negatively affect the results on the EmoInt dataset.

Unexpectedly, retrofitting does not help in all settings in our post-processing pipeline except for

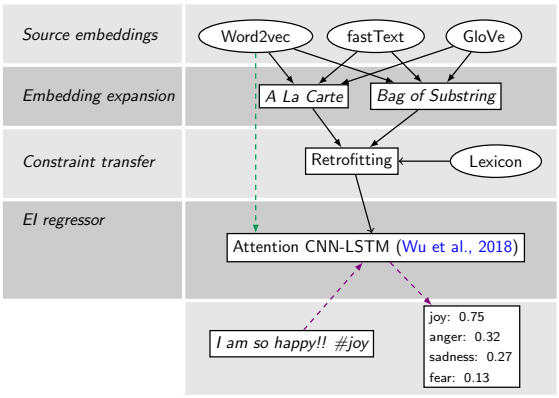

Figure 6: Experimental Setup. The green arrow from Word2vec to the regressor unit shows the information flow in the baseline. The black solid arrows show the different experimental settings. The purple dashed arrows at the bottom show the prediction phase.

fastText embeddings. We assume that is a consequence of using a too small lexicon for retrofitting, and the method would improve the embeddings if sentiment or emotion lexicons would be used instead. However, this needs further investigation.

5 Conclusion & Future Work

With this paper, we presented the first analysis of the distribution of degree adverbs and negations on Twitter in the context of emotions. In addition, we proposed a pipeline with different modules to expand embeddings particularly for emotion phrases. Our evaluation shows substantial differences based on the combination of input embeddings and the postprocessing method. Our pipeline improves the results obtained while evaluating the downstream task of emotion intensity prediction on our dataset. Finally, we contribute a novel emotion phrase lexicon of high precision.

For future work we propose to analyze other baseline approaches, particularly learning a composition function over pairs of adjectives with degree adverbs. The modifiers could be considered as functions over adjectives and would be represented as matrices.

Another further improvement of this work would be to expand this analysis to verbal and nominal expressions of emotion, which we hypothesize as also being frequent. In order to obtain meaningful representations for the phrases we focus on, another natural next step is expanding the postprocessing pipeline and including a comparison to other adaptation methods such as counterfitting

	joy		sadness		anger		fear		*average*	
	T	EI	T	EI	T	EI	T	EI	T	EI
G	.20	.60	.21	.59	.24	.60	.27	.61	.23	.60
G+ALC	.23	.61	.31	.63	.33	.62	.35	.62	.30	.63
G+BoS	.24	.58	.30	.60	.34	.59	.36	.57	.31	.59
G+ALC+RF	.19	.60	.21	.61	.26	.63	.28	.61	.24	.61
G+BoS+RF	.19	.62	.21	.60	.28	.62	.25	.61	.23	.61
W2V	.16	.60	.12	.59	.19	.60	.23	.63	.18	.62
W2V+ALC	.20	.60	.24	.64	.28	.65	.28	.64	.25	.63
W2V+BoS	.20	.61	.23	.64	.28	.66	.29	.60	.25	.64
W2V+ALC+RF	.21	.60	.25	.54	.28	.69	.28	.64	.26	.62
W2V+BoS+RF	.16	.60	.12	.61	.24	.67	.20	.60	.18	.63
FT	.16	.58	.14	.53	.21	.65	.22	.60	.18	.61
FT+ALC	.16	.59	.14	.52	.21	.59	.23	.62	.19	.59
FT+BoS	.16	.60	.14	.59	.22	.63	.23	.61	.18	.62
FT+ALC+RF	.18	.54	.16	.62	.22	.64	.25	.59	.20	.60
FT+ BoS+RF	.16	.60	.14	.57	.22	.62	.21	.57	.18	.63

Table 3: Evaluation: Spearman's rank correlation between predicted emotion intensity scores and annotated scores on our dataset (T) or the EmoInt dataset (EI). We report results only for the 4 emotions annotated in the EmoInt data.

(Mrksić et al., 2016). Presumably, this will also generate additional insights into the aspect that we were only able to show a limited improvement based on retrofitting.

Given the recent advances in representing contextualized word embeddings as functions computing dynamically the embeddings for words given their context, we hypothesize and intend to further verify that these embeddings would be a better choice for input to systems that predict intensity scores. It would be interesting to compare models such as word embeddings from language models (Elmo) (Peters et al., 2018), bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018), and generative pre-training OpenAI (GPT) (Radford et al., 2019) to the ones we already discussed, since the contextualized embeddings assign a different vector for a word in each given context. These approaches presumably produce a different vector for "happy" in the context of "not" than in the content of "very" or "completely".

Lastly, we plan to also adjust the lexica created such that it covers more domains, sources, and languages.

Acknowledgements

This research has been funded by the German Research Council (DFG), projects SEAT (Structured Multi-Domain Emotion Analysis from Text, KL

2869/1-1). We thank Jeremy Barnes, Evgeny Kim, Sean Papay, Sebastian Padó and Enrica Troiano for fruitful discussions and the reviewers for the helpful comments.

References

Laura Aina, Raffaella Bernardi, and Raquel Fernández. 2018. A distributional study of negated adjectives and antonyms. In *Proceedings of CLiC-it 2018 the 5th Italian Conference on Computational Linguistics*.

Jorge Carrillo-de Albornoz and Laura Plaza. 2013. An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology*, 64(8):1618–1633.

Angeliki Athanasiadou. 2007. On the subjectivity of intensifiers. *Language sciences*, 29(4):554–565.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Isaac Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59. University of Antwerp.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eduard Dragut and Christiane Fellbaum. 2014. The role of adverbs in sentiment analysis. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 38–41. Association for Computational Linguistics.

E. Ehrlich. 1980. *Oxford American Dictionary*. Oxford University Press.

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

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2016a. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 43–52. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016b. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016c. Sentiment composition of words with opposing polarities. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1102–1108. Association for Computational Linguistics.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Saif Mohammad and Felipe Bravo-Marquez. 2017a. 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. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017b. 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. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17. Association for Computational Linguistics.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Nikola Mrksić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148. Association for Computational Linguistics.

Maria Napoli and Miriam Ravetto. 2017. New insights on intensification and intensifiers. *Exploring Intensification: Synchronic, diachronic and cross-linguistic perspectives*, 189:1.

Terttu Nevalainen. 2008. Social variation in intensifier use: constraint on-ly adverbialization in the past? *English Language & Linguistics*, 12(2):289–315.

Terttu Nevalainen and Matti Rissanen. 2002. Fairly pretty or pretty fair? on the development and grammaticalization of english downtoners. *Language Sciences*, 24(3-4):359–380.

Carita Paradis. 1997. *Degree modifiers of adjectives in spoken British English*, volume 92 of *Lund Studies in English*. Lund University Press.

Carita Paradis. 2000. It's well weird: Degree modifiers of adjectives revisited: The nineties. *Language and computers*, 30:147–160.

Carita Paradis. 2001. Adjectives and boundedness.

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. Association for Computational Linguistics.

Paloma Núñez Pertejo and Ignacio M Palacios Martínez. 2014. That's absolutely crap, totally rubbish: The use of the intensifiers absolutely and totally in the spoken language of british adults and teenagers. *Functions of Language*, 21(2):210–237.

Hans Peters. 1994. *Degree Adverbs in Early Modern English*, volume 13. De Gruyter Mouton.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1(3-31):4.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734.

R. Quirk. 1985. *A Comprehensive grammar of the English language*. General Grammar Series. Longman.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. 2015. Negation scope detection for twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108. Association for Computational Linguistics.

A. Stevenson and C.A. Lindberg. 2010. *New Oxford American Dictionary, Third Edition*. OUP USA.

Florian Strohm and Roman Klinger. 2018. An empirical analysis of the role of amplifiers, downtoners, and negations in emotion classification in microblogs. In *The 5th IEEE International Conference on Data Science and Advanced Analytics, Special Track on Sentiment, Emotion, and Credibility of Information in Social Data*, DSAA, Turin, Italy. IEEE.

Sali A Tagliamonte. 2006. "So cool, right?": Canadian english entering the 21st century. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 51(2-3):309–331.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Zhigang Yuan, Sixing Wu, and Yongfeng Huang. 2018. Thu_ngn at semeval-2018 task 1: Fine-grained tweet sentiment intensity analysis with attention cnn-lstm. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 186–192. Association for Computational Linguistics.

Jinman Zhao, Sidharth Mudgal, and Yingyu Liang. 2018. Generalizing word embeddings using bag of subwords. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 601–606. Association for Computational Linguistics.

Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 304–313. Association for Computational Linguistics.