Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and **Roman Klinger**[*]
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{firstname,lastname}@ims.uni-stuttgart.de

Abstract

There is a rich variety of data sets for sentiment analysis (*viz.*, polarity and subjectivity classification). For the more challenging task of detecting discrete emotions following the definitions of Ekman and Plutchik, however, there are much fewer data sets, and notably no resources for the social media domain. This paper contributes to closing this gap by extending the SemEval 2016 stance and sentiment dataset with emotion annotation. We (a) analyse annotation reliability and annotation merging; (b) investigate the relation between emotion annotation and the other annotation layers (stance, sentiment); (c) report modelling results as a baseline for future work.

1 Introduction

Emotion recognition is a research area in natural language processing concerned with associating words, phrases or documents with predefined emotions from psychological models. *Discrete emotion recognition* assigns categorial emotions (Ekman, 1999; Plutchik, 2001), namely *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise* und *Trust*. Compared to the very active area of sentiment analysis, whose goal is to recognize the polarity of text (*e. g.*, positive, negative, neutral, mixed), few resources are available for discrete emotion analysis.

Emotion analysis has been applied to several domains, including tales (Alm et al., 2005), blogs (Aman and Szpakowicz, 2007) and microblogs (Dodds et al., 2011). The latter in particular provides a major data source in the form of user messages from platforms such as Twitter (Costa et al., 2014) which contain semi-structured information (hashtags, emoticons, emojis) that can be used as weak supervision for training classifiers (Suttles and Ide, 2013). The classifier then learns the association of all other words in the message with the "self-labeled" emotion (Wang et al., 2012).

While this approach provides a practically feasible approximation of emotions, there is no publicly available, manually vetted data set for Twitter emotions that would support accurate and comparable evaluations. In addition, it has been shown that distant annotation is conceptually different from manual annotation for sentiment and emotion (Purver and Battersby, 2012).

With this paper, we contribute manual emotion annotation for a publicly available Twitter data set. We annotate the SemEval 2016 Stance Data set (Mohammad et al., 2016) which provides sentiment and stance information and is popular in the research community (Augenstein et al., 2016; Wei et al., 2016; Dias and Becker, 2016; Ebrahimi et al., 2016). It therefore enables further research on the relations between sentiment, emotions, and stances. For instance, if the distribution of subclasses of positive or negative emotions is different for *against* and *in-favor*, emotion-based features could contribute to stance detection.

An additional feature of our resource is that we do not only provide a "majority annotation" as is usual. We do define a well-performing aggregated annotation, but additionally provide the *individual labels* of each of our six annotators. This enables further research on differences in the perception of emotions.

2 Background and Related Work

For a review of the fundaments of emotion and sentiment and the differences between these concepts, we refer the reader to Munezero et al. (2014).

[*] We thank Marcus Hepting, Chris Krauter, Jonas Vogelsang, Gisela Kollotzek for annotation and discussion.

Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 13–23
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

Name	Granularity	Annotation	Size	Topic	Source
STS-test	tweet	1	498	General	Go et al. (2009)
SemEval 2013	tweet	2	15,196	General	Nakov et al. (2013)
Healthcare Reform	tweet	2	2,516	Politics	Speriosu et al. (2011)
Obama-McCain Debate	tweet	3	3,238	Politics	Shamma et al. (2009)
Dialogue Earth-WA	tweet	4	4,490	Weather	Cavender-Bares (2011)
Dialogue Earth-WB	tweet	4	8,850	Weather	Busch (2011)
Dialogue Earth-GASP	tweet	4	12,770	Gas prices	Busch (2012)
STS-GOLD	entity/tweet	5	2,205	General	Hassan Saif and Alani (2013)
SemEval 2016	topics/tweets	6	4,870	5 topics	Mohammad et al. (2016)
Sentiment Strength	tweet	7	4,242	General	Thelwall et al. (2012)
ISEAR	descriptions	8	7,666	Emotional Events	Scherer and Wallbott (1997)
Tales	sentences	9	1,580	Grim's Fairytales	Alm et al. (2005)
Blogs	blogs	10	173	General	Aman and Szpakowicz (2007)
SemEval 2017	headlines	11	1,250	General	Strapparava and Mihalcea (2007)
WASSA EmoInt 2017	tweets	12	7,102	General	Mohammad and Bravo-Marquez (2017)
Electoral Tweets	tweets	13	965	Elections	Mohammad et al. (2015)

Table 1: A selection of resources for sentiment analysis (on Twitter, 1–7) and emotion analysis (in general, 8–12). Annotation refers to the following annotation schemes: [1] positive-negative, [2] positive-negative-neutral, [3] positive-negative-mixed-other, [4] positive-negative-netural-unrelated-can't tell, [5] positive-negative-neutral-mixed-other, [6] for-against, [7] positive and negative strength (range), [8] joy, fear, anger, sadness, disgust, shame, guilt, [9] angry, disgusted, fearful, happy, sad, positively surprised, negatively surprised, [10] happiness, sadness, anger, disgust, surprise, fear, mixed, [11] anger, disgust, fear, joy, sadness, surprise, [12] anger, fear, joy, sadness, [13] positive, negative, mixed, intensity, trust, fear, surprise, disgust, anger, anticipation, joy, roles, style, purpose (number denotes subset in corpus with emotion annotations)

For sentiment analysis, a large number of annotated data sets exists. These include review texts from different domains, for instance from Amazon and other shopping sites (Hu and Liu, 2004; Ding et al., 2008; Toprak et al., 2010; Lakkaraju et al., 2011), restaurants (Ganu et al., 2009), news articles (Wiebe et al., 2005), blogs (Kessler et al., 2010), as well as microposts on Twitter. For the latter, shown in the upper half of Table 1, there are general corpora (Nakov et al., 2013; Spina et al., 2012; Thelwall et al., 2012) as well as ones focused on very specific subdomains, for instance on Obama-McCain Debates (Shamma et al., 2009), Health Care Reforms (Speriosu et al., 2011). A popular example for a manually annotated corpus for sentiment, which includes stance annotation for a set of topics is the SemEval 2016 data set (Mohammad et al., 2016).

For emotion analysis, the set of annotated resources is smaller (compare the lower half of Table 1). A very early resource is the ISEAR data set (Scherer and Wallbott, 1997) which contains descriptions of emotional events. While motivated by psychological research, it was later repurposed for computational research. The first data set developed specifically for computational research was the tales corpus by Alm et al. (2005). Aman and Szpakowicz (2007) published a corpus of blog posts. In the context of SemEval, Strapparava and Mihalcea (2007) annotated news headlines.

A notable gap is the unavailability of a publicly available set of microposts (*e. g.*, tweets) with emotion labels. To the best of our knowledge, there are only three previous approaches to labeling tweets with discrete emotion labels. One is the recent data set on for emotion intensity estimation, a shared task aiming at the development of a regression model. The goal is not to predict the emotion class, but a distribution over their intensities, and the set of emotions is limited to *fear*, *sadness*, *anger*, and *joy* (Mohammad and Bravo-Marquez, 2017).

Most similar to our work is a study by Roberts et al. (2012) which annotated 7,000 tweets manually for 7 emotions (anger, disgust, fear, joy, love, sadness and surprise). They chose 14 topics which they believe should elicit emotional tweets and collect hashtags to help identify tweets that are on these topics. After several iterations, the annotators reached $\kappa = 0.67$ inter-annotator agreement on 500 tweets. Unfortunately, the data appear not to be available any more. An additional limitation of that dataset was that 5,000 of the 7,000 tweets were annotated by one annotator only. In contrast, we provide several annotations for each tweet.

	Label count for threshold t				
Emotion	0.0	0.33	0.5	0.66	0.99
Anger	2,902	2,238	1,388	1,315	578
Anticipation	2,700	1,656	739	677	199
Disgust	2,183	1,199	440	404	106
Fear	1,840	895	274	246	68
Joy	2,067	1,384	815	764	402
Sadness	2,644	1,389	414	343	78
Surprise	1,108	489	177	156	33
Trust	1,713	984	520	487	213

Table 2: Corpus Statistics. The threshold t measures that a fraction of more than t annotators labeled the respective emotion (*e. g.*, t=0.0: at least one annotator t=0.99: all annotators). Overall number of tweets: 4,868.

	Cohen's κ	
Emotion	Min	Max
Anger	0.28	0.49
Anticipation	0.11	0.39
Disgust	0.06	0.30
Fear	0.08	0.25
Joy	0.30	0.52
Sadness	0.04	0.30
Surprise	0.09	0.33
Trust	0.29	0.57

Table 3: Kappa Statistics for all pairs of annotators.

Mohammad et al. (2015) annotated electoral tweets for sentiment, intensity, semantic roles, style, purpose and emotions. This is the only available corpus similar to our work we are aware of. However, the focus of this work was not emotion annotation in contrast to ours. In addition, we publish the data of all annotators.

3 Corpus Annotation and Analysis

3.1 Annotation Procedure

As motivated above, we re-annotate the extended SemEval 2016 Stance Data set (Mohammad et al., 2016) which consists of 4,870 tweets (a subset of which was used in the SemEval competition). For a discussion of the differences of these data sets, we refer to Mohammad et al. (2017). We omit two tweets with special characters, which leads to an overall set of 4,868 tweets used in our corpus.[1]

We frame annotation as a multi-label classification task at the tweet level. The tweets were annotated by a group of six independent annotators, with a minimum number of three annotations for each tweet (696 tweets were labeled by 6 annotators, 703 by 5 annotators, 2,776 by 4 annotators and 693 by 3 annotators). All annotators were undergraduate students of media computer science and between the age of 20 and 30. Only one annotator is female. All students are German native speakers and have college-level proficiency in English. To train the annotators on the task, we performed two training iterations based on 50 randomly selected tweets from the SemEval 2016 Task 4 corpus (Nakov et al., 2016). After each iteration, we discussed annotation differences (informally) in face-to-face meetings.

For the final annotation, tweets were presented to the annotators in a web interface which paired a tweet with a set of binary check boxes, one for each emotion. Taggers could annotate any set of emotions. Each annotator was assigned with 5/7 of the corpus with equally-sized overlap of instances based on an offset shift. Not all annotators finished their task.[2]

3.2 Emotion Annotation Reliability and Aggregated Annotation

Our annotation represents a middle ground between traditional linguistic "expert" annotation and crowdsourcing: We assume that intuitions about emotions diverge more than for linguistic structures. At the same time, we feel that there is information in the individual annotations beyond the simple "majority vote" computed by most crowdsourcing studies. In this section, we analyse the annotations intrinsically; a modelling-based evaluation follows in Section 5.

Our first analysis, shown in Table 2, compares annotation strata with different agreement. For example, the column labeled 0.0 lists the frequencies of emotion labels assigned by at least one annotator, a *high recall* annotation. In contrast, the column labeled 0.99 lists frequencies for emotion labels that all annotators agreed on. This represents a *high*

[1] Our annotations and original tweets are available at `http://www.ims.uni-stuttgart.de/data/ssec` and `http://alt.qcri.org/semeval2016/task6/data/uploads/stancedataset.zip`, see also `http://alt.qcri.org/semeval2016/task6`.

[2] Initially, we recruited seven annotators. One annotator dropped out; we do not publish their data.

		Emotions								Sentiment			Stance		
		Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Positive	Negative	Neutral	In Favor	Against	None
Emotion	Anger	**2902**	1437	1983	1339	774	2065	711	640	275	2534	93	630	1628	644
	Anticipation	0.55	**2700**	1016	1029	1330	1369	482	1234	1094	1445	161	772	1291	637
	Disgust	19.05	0.52	**2183**	1024	512	1628	526	404	126	2008	49	429	1291	463
	Fear	2.51	1.03	2.02	**1840**	466	1445	407	497	306	1445	89	448	982	410
	Joy	0.19	1.88	0.22	0.30	**2067**	682	438	1101	1206	750	111	596	952	519
	Sadness	5.91	0.72	4.82	5.58	0.21	**2644**	664	613	345	2171	128	604	1429	611
	Surprise	1.28	0.54	1.15	0.94	0.86	1.34	**1108**	222	219	801	88	257	521	330
	Trust	0.24	2.97	0.24	0.55	4.08	0.31	0.38	**1713**	1082	558	73	500	860	353
Sent.	Positive	0.06	2.75	0.06	0.30	10.94	0.13	0.46	10.53	**1524**	0	0	485	673	366
	Negative	20.3	0.42	18.61	3.32	0.13	7.27	1.79	0.13	0.0	**3032**	0	622	1665	745
	Neutral	0.26	0.85	0.21	0.64	0.73	0.56	1.36	0.54	0.0	0.0	**312**	97	71	144
Stance	In Favor	0.67	1.61	0.60	0.97	1.46	0.80	0.90	1.44	1.70	0.56	1.41	**1204**	0	0
	Against	1.94	0.86	2.03	1.28	0.79	1.49	0.88	1.05	0.73	1.79	0.28	0.0	**2409**	0
	None	0.63	0.77	0.64	0.74	0.94	0.74	1.30	0.65	0.87	0.85	2.66	0.0	0.0	**1255**

Table 4: Tweet Counts (above diagonal) and odds ratio (below diagonal) for cooccurring annotations for all classes in the corpus (emotions based on aggregated annotation, t=0.0).

precision annotation. The other levels represent intermediate precision-recall trade-offs.

These numbers confirm that emotion labeling is a somewhat subjective task: only a small subset of the emotions labeled by at least one annotator (t=0.0) is labeled by most (t=0.66) or all of them (t=0.99). Interestingly, the exact percentage varies substantially by emotion, between 2 % for *sadness* and 20 % for *anger*.

Many of these disagreements stem from tweets that are genuinely difficult to categorize emotionally, like

> *That moment when Canadians realised global warming doesn't equal a tropical vacation*

for which one annotator chose *anger* and *sadness*, while one annotator chose surprise. Arguably, both annotations capture aspects of the meaning. Similarly, the tweet

> *2 pretty sisters are dancing with cancered kid*

(a reference to an online video) is marked as *fear* and *sadness* by one annotator and with *joy* and *sadness* by another. Naturally, not all differences arise from justified annotations. For instance the tweet

> *#BIBLE = Big Irrelevant Book of Lies and Exaggerations*

has been labeled by two annotators with the emotion *trust*, presumably because of the word *bible*. This appears to be a classical oversight error, where the tweet is labeled on the basis of the first spotted keyword, without substantially studying its content.

To quantify these observations, we follow general practice and compute a chance-corrected measure of inter-annotator agreement. Table 3 shows the minimum and maximum Cohen's κ values for pairs of annotators, computed on the intersection of instances annotated by either annotator within each pair. We obtain relatively high κ values of *anger*, *joy*, and *trust*, but lower values for the other emotions.

These small κ values could be interpreted as indicators of problems with reliability. However, κ is notoriously difficult to interpret, and a number of studies have pointed out the influence of marginal frequencies (Cicchetti and Feinstein, 1990): In the presence of skewed marginals (and most of our emotion labels are quite rare, *cf.* Table 2), the expected agreement (referred to as $P(E)$ in contrast to $P(A)$ for the empirical agreement) is quite high. This makes it hard to obtain high κ values; thus, low κ values do not necessarily indicate unreliable annotation.

To avoid these methodological problems, we assess the usefulness of our annotation extrinsically by comparing the performance of computational models for different values of t. In a nutshell, these experiments will show best results t=0.0, *i. e.*, the

		Emotions								Sentiment			Stance		
		Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Positive	Negative	Neutral	In Favor	Against	None
Emotion	Anger	**1388**	53	334	87	37	195	63	12	28	1353	7	272	840	276
	Anticipation	0.16	**739**	16	42	218	14	2	182	445	253	41	258	333	148
	Disgust	10.09	0.19	**440**	39	11	72	26	2	1	439	0	67	289	84
	Fear	1.18	1.01	1.74	**274**	4	58	9	13	26	241	7	83	116	75
	Joy	0.10	2.48	0.12	0.07	**815**	7	9	196	658	142	15	263	304	248
	Sadness	2.43	0.18	2.34	3.20	0.08	**414**	14	3	28	377	9	102	216	96
	Surprise	1.40	0.06	1.78	0.89	0.26	0.92	**177**	0	16	145	16	46	76	55
	Trust	0.05	3.66	0.03	0.40	3.64	0.06	0.0	**520**	462	43	15	142	337	41
Sent.	Positive	0.03	4.28	0.0	0.22	15.42	0.14	0.21	24.65	**1524**	0	0	485	673	366
	Negative	41.47	0.25	310.67	4.72	0.08	6.90	2.83	0.04	0.0	**3032**	0	622	1665	745
	Neutral	0.05	0.84	0.0	0.37	0.24	0.30	1.48	0.41	0.0	0.0	**312**	97	71	144
Stance	In Favor	0.67	1.80	0.52	1.35	1.58	0.99	1.07	1.16	1.70	0.56	1.41	**1204**	0	0
	Against	1.87	0.81	2.08	0.74	0.55	1.12	0.76	2.02	0.73	1.79	0.28	0.0	**2409**	0
	None	0.63	0.68	0.66	1.09	1.32	0.86	1.31	0.22	0.87	0.85	2.66	0.0	0.0	**1255**

Table 5: Tweet Counts (above diagonal) and odds ratio (below diagonal) for cooccurring annotations for all classes in the corpus (emotions based on majority annotation, t=0.5).

high-recall annotation (see Section 5 for details). We therefore define t=0.0 as our *aggregated annotation*. For comparison, we also consider t=0.5, which corresponds to the *majority annotation* as generally adopted in crowdsourcing studies.

3.3 Distribution of Emotions

As shown in Table 2, nearly 60 % of the overall tweet set are annotated with *anger* by at least one annotator. This is the predominant emotion class, followed by *anticipation* and *sadness*. This distribution is comparably uncommon and originates from the selection of tweets in SemEval as a stance data set. However, while *anger* clearly dominates in the aggregated annotation, its predominance weakens for the more precision-oriented data sets. For t=0.99, *joy* becomes the second most frequent emotion. In uniform samples from Twitter, joy typically dominates the distribution of emotions (Klinger, 2017). It remains a question for future work how to reconciliate these observations.

3.4 Emotion vs. other Annotation Layers

Table 4 shows the number of cooccurring label pairs (above the diagonal) and the odds ratios (below the diagonal) for emotion, stance, and sentiment annotations on the whole corpus for our aggregated annotation (t=0.0). Odds ratio is

$$\text{R(A:B)} = \frac{P(A)(1 - P(B))}{P(B)(1 - P(A))} ,$$

where $P(A)$ is the probability that both labels (at row and column in the table) hold for a tweet and $P(B)$ is the probability that only one holds. A ratio of x means that the joint labeling is x times more likely than the independent labeling. Table 5 shows the same numbers for the majority annotation, t=0.5.

We first analyze the relationship between emotions and sentiment polarity in Table 4. For many emotions, the polarity is as expected: *Joy* and *trust* occur predominantly with positive sentiment, and *anger, disgust*, *fear* and *sadness* with negative sentiment. The emotions *anticipation* and *surprise* are, in comparison, most balanced between polarities, however with a majority for positive sentiment in anticipation and a negative sentiment for surprise. For most emotions there is also a non-negligible number of tweets with the sentiment opposite to a common expectation. For example, *anger* occurs 28 times with positive sentiment, mainly tweets which call for (positive) change regarding a controversial topic, for instance

> *Lets take back our country! Whos with me? No more Democrats!2016*
>
> *Why criticise religions? If a path is not your own. Don't be pretentious. And get down from your throne.*

Conversely, more than 15 % of the *joy* tweets carry negative sentiment. These are often cases in which either the emotion annotator or the sentiment annotator assumed some non-literal meaning to be associated with the text (mainly irony), for instance

> *Global Warming! Global Warming! Global Warming! Oh wait, it's summer.*
>
> *I love the smell of Hillary in the morning. It smells like Republican Victory.*

Disgust occurs almost exclusively with negative sentiment.

For the majority annotation (Table 5), the number of annotations is smaller. However, the average size of the odds ratios increase (from 1.96 for t=0.0 to 5.39 for t=0.5).

A drastic example is *disgust* in combination with negative sentiment, the predominant combination. *Disgust* is only labeled once with positive sentiment in the t=0.5 annotation:

> *#WeNeedFeminism because #NoMeansNo it doesnt mean yes, it doesnt mean try harder!*

Similarly, the odds ratio for the combination *anger* and negative sentiment nearly doubles from 20.3 for t=0.0 to 41.47 for t=0.5. These numbers are an effect of the majority annotation having a higher precision in contrast to more "noisy" aggregation of all annotations (t=0.0).

Regarding the relationship between emotions and stance, most odds ratios are relatively close to 1, indicating the absence of very strong correlations. Nevertheless, the "Against" stance is associated with a number of negative emotions (*anger, disgust, sadness*, the "In Favor" stance with *joy, trust,* and *anticipation*, and "None" with an absence of all emotions except *surprise*.

4 Models

We apply six standard models to provide baseline results for our corpus: Maximum Entropy (MaxEnt), Support Vector Machines (SVM), a Long-Short Term Memory Network (LSTM), a Bidirectional LSTM (Bi-LSTM), and a Convolutional Neural Network (CNN).

MaxEnt and **SVM** classify each tweet separately based on a bag-of-words. For the first, the linear separator is estimated based on log-likelihood optimization with an L2 prior. For the second, the optimization follows a max-margin strategy.

LSTM (Hochreiter and Schmidhuber, 1997) is a recurrent neural network architecture which includes a memory state capable of learning long distance dependencies. In various forms, they have proven useful for text classification tasks (Tai et al., 2015; Tang et al., 2016). We implement a standard LSTM which has an embedding layer that maps the input (padded when needed) to a 300 dimensional vector. These vectors then pass to a 175 dimensional LSTM layer. We feed the final hidden state to a fully-connected 50-dimensional dense layer and use sigmoid to gate our 8 output neurons. As a regularizer, we use a dropout (Srivastava et al., 2014) of 0.5 before the LSTM layer.

Bi-LSTM has the same architecture as the normal LSTM, but includes an additional layer with a reverse direction. This approach has produced state-of-the-art results for POS-tagging (Plank et al., 2016), dependency parsing (Kiperwasser and Goldberg, 2016) and text classification (Zhou et al., 2016), among others. We use the same parameters as the LSTM, but concatenate the two hidden layers before passing them to the dense layer.

CNN has proven remarkably effective for text classification (Kim, 2014; dos Santos and Gatti, 2014; Flekova and Gurevych, 2016) . We train a simple one-layer CNN with one convolutional layer on top of pre-trained word embeddings, following Kim (2014). The first layer is an embeddings layer that maps the input of length n (padded when needed) to an n x 300 dimensional matrix. The embedding matrix is then convoluted with filter sizes of 2, 3, and 4, followed by a pooling layer of length 2. This is then fed to a fully connected dense layer with ReLu activations and finally to the 8 output neurons, which are gated with the sigmoid function. We again use dropout (0.5), this time before and after the convolutional layers.

For all neural models, we initialize our word representations with the skip-gram algorithm with negative sampling (Mikolov et al., 2013), trained on nearly 8 million tokens taken from tweets collected using various hashtags. We create 300-dimensional vectors with window size 5, 15 negative samples and run 5 iterations. For OOV words, we use a vector initialized randomly between -0.25 and 0.25 to approximate the variance of the pretrained vectors. We train our models using ADAM (Kingma and Ba, 2015) and a minibatch size of 32. We set 10 % of

	Results for Threshold $t = 0.0$ for standard models														
	Linear						Neural								
	MaxEnt			SVM			LSTM			Bi-LSTM			CNN		
Emotion	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	76	72	74	76	69	72	76 (1.7)	77 (5.3)	76 (1.9)	77 (0.8)	77 (2.7)	**77** (1.3)	77 (0.8)	77 (2.7)	**77** (1.3)
Anticipation	72	61	66	70	60	64	68 (1.8)	68 (8.9)	67 (3.5)	70 (1.2)	66 (3.6)	**68** (1.6)	68 (1.2)	60 (0.8)	64 (0.5)
Disgust	62	47	54	59	53	56	64 (3.2)	68 (8.7)	**65** (2.5)	61 (1.4)	64 (4.6)	63 (1.7)	62 (0.6)	61 (3.9)	62 (1.9)
Fear	57	31	40	55	40	46	51 (3.5)	48 (8.5)	**49** (4.6)	58 (1.6)	43 (6.3)	**49** (3.8)	53 (1.7)	46 (6.2)	**49** (3.9)
Joy	55	50	52	52	52	52	56 (5.9)	41 (8.3)	46 (4.8)	54 (2.9)	59 (10.5)	**56** (4.8)	54 (1.7)	56 (5.6)	55 (2.3)
Sadness	65	65	65	64	60	62	60 (2.5)	77 (11.1)	**67** (3.9)	62 (0.6)	72 (7.5)	**67** (3.2)	63 (0.9)	72 (0.3)	**67** (0.5)
Surprise	62	15	24	46	22	30	40 (4.4)	17 (10.4)	21 (8.7)	42 (2.9)	20 (3.2)	27 (2.5)	36 (3.7)	24 (6.3)	**28** (5.0)
Trust	62	38	47	57	45	50	57 (6.1)	49 (12.3)	**51** (5.9)	59 (2.5)	44 (4.1)	50 (2.5)	53 (0.6)	49 (6.6)	50 (3.3)
Micro-Avg.	66	52	58	63	53	58	62 (0.9)	60 (1.9)	61 (0.7)	64 (0.3)	60 (2.4)	**62** (1.2)	62 (0.6)	59 (2.0)	60 (1.0)

Table 6: Results of linear and neural models for labels from the aggregated annotation (t=0.0). For the neural models, we report the average of five runs and standard deviation in brackets. Best F_1 for each emotion shown in boldface.

the training data aside to tune the hyperparameters for each model (hidden dimension size, dropout rate, and number of training epochs).

5 Results

Table 6 shows the results for our canonical annotation aggregation with t=0.0 (aggregated annotation) for our models. The two linear classifiers (trained as MaxEnt and SVM) show comparable results, with an overall micro-average F_1 of 58 %. All neural network approaches show a higher performance of at least 2 percentage points (3 pp for LSTM, 4 pp for Bi-LSTM, 2 pp for CNN). Bi-LSTM also obtains the best F-Score for 5 of the 8 emotions (4 out of 8 for LSTM and CNN). We conclude that the Bi-LSTM shows the best results of all our models. Our discussion focuses on this model.

The performance clearly differs between emotion classes. Recall from Section 3.2 that *anger, joy* and *trust* showed much higher agreement numbers than the other annotations. There is however just a mild correlation between reliability and modeling performance. *Anger* is indeed modelled very well: it shows the best prediction performance with a similar precision and recall on all models. We ascribe this to it being the most frequent emotion class. In contrast, *joy* and *trust* show only middling performance, while we see relatively good results for *anticipation* and *sadness* even though there was considerable disagreement between annotators. We find the overall worst results for *surprise*. This is not surprising, *surprise* being a scarce label with also very low agreement. This might point towards underlying problems in the definition of *surprise* as an emotion. Some authors have split this class into positive and negative surprise in an attempt to avoid this (Alm et al., 2005).

We finally come to our justification for choosing t=0.0 as our aggregated annotation. Table 7 shows results for the best model (Bi-LSTM) on the datasets for different thresholds. We see a clear downward monotone trend: The higher the threshold, the lower the F_1 measures. We obtain the best results, both for individual emotions and at the average level, for t=0.0. This is at least partially counterintuitive – we would have expected a dataset with "more consensual" annotation to yield better models – or at least models with higher precision. This is not the case. Our interpretation is that frequency effects outweigh any other considerations: As Table 2 shows, the amount of labeled data points drops sharply with higher thresholds: even between t=0.0 and t=0.33, on average half of the labels are lost. This interpretation is supported by the behavior of the individual emotions: for emotions where the data sets shrink gradually (*anger, joy*), performance drops gradually, while it dips sharply for emotions where the data sets shrink fast (*disgust, fear*). Somewhat surprisingly, therefore, we conclude that t=0.0 appears to be the

	Results of BiLSTM for different voting thresholds t														
	0.0			0.33			0.5			0.66			0.99		
Emotion	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	77	76	77	64	71	68	52	45	48	47	51	49	34	14	20
	(1.3)	(4.8)	(1.9)	(1.7)	(3.8)	(1.5)	(0.6)	(7.8)	(4.8)	(1.5)	(6.7)	(2.6)	(5.2)	(2.6)	(2.4)
Anticipation	70	66	68	60	43	50	42	23	29	37	20	25	11	12	11
	(1.2)	(3.6)	(1.6)	(2.3)	(5.6)	(3.4)	(5.9)	(4.4)	(2.8)	(4.1)	(7.6)	(6.0)	(3.3)	(2.9)	(1.9)
Disgust	61	64	63	48	38	42	34	13	18	24	8	11	11	2	3
	(1.4)	(4.6)	(1.7)	(1.5)	(4.4)	(2.6)	(4.9)	(2.7)	(3.5)	(6.2)	(2.7)	(3.7)	(10.7)	(2.0)	(3.4)
Fear	58	43	49	34	22	26	18	15	13	11	14	11	1	6	1
	(1.6)	(6.3)	(3.8)	(3.2)	(5.9)	(4.6)	(8.1)	(10.5)	(5.3)	(5.0)	(10.5)	(7.9)	(1.3)	(11.7)	(2.3)
Joy	54	59	56	56	41	47	53	37	43	54	34	41	64	27	35
	(2.9)	(10.5)	(4.8)	(2.8)	(6.3)	(3.6)	(4.3)	(3.6)	(1.6)	(7.1)	(4.2)	(2.1)	(14.9)	(9.6)	(6.8)
Sadness	62	72	67	42	47	44	16	24	19	15	19	16	3	6	4
	(0.6)	(7.5)	(3.2)	(1.4)	(6.2)	(2.1)	(2.1)	(6.0)	(2.0)	(2.3)	(7.6)	(3.0)	(2.0)	(2.9)	(1.9)
Surprise	42	20	27	31	20	23	12	20	13	12	12	12	0	0	0
	(2.9)	(3.2)	(2.5)	(6.8)	(7.5)	(3.2)	(2.3)	(8.9)	(2.1)	(1.3)	(2.6)	(1.7)	(0.0)	(0.0)	(0.0)
Trust	59	44	50	66	31	42	60	24	34	59	23	33	35	14	18
	(2.5)	(4.1)	(2.5)	(3.4)	(2.7)	(2.3)	(4.6)	(7.1)	(7.1)	(3.5)	(6.8)	(6.8)	(7.4)	(11.2)	(9.7)
Micro-Avg.	64	60	62	53	44	48	38	30	33	38	29	33	21	14	17
	(0.3)	(2.4)	(1.2)	(1.8)	(1.8)	(0.6)	(2.2)	(3.3)	(2.4)	(1.8)	(4.1)	(2.9)	(4.2)	(3.1)	(3.2)

Table 7: Results of the BiLSTM for different voting thresholds. We report average results for each emotion over 5 runs (standard deviations are included in parenthesis).

most useful datasets from a computational modeling perspective.

In terms of how to deal with diverging annotations, we believe that this result bolsters our general approach to pay attention to individual annotators' labels rather than just majority votes: if the individual labels were predominantly noisy, we would not expect to see relatively high F_1 scores.

6 Conclusion and Future Work

With this paper, we publish the first manual emotion annotation for a publicly available micropost corpus. The resource we chose to annotate already provides stance and sentiment information. We analyzed the relationships among emotion classes and between emotions and the other annotation layers.

In addition to the data set, we implemented well-known standard models which are established for sentiment and polarity prediction for emotion classification. The Bi-LSTM model outperforms all other approaches by up to 4 points F_1 on average compared to linear classifiers.

Inter-annotator analysis showed a limited agreement between the annotators – the task is, at least to some degree, driven by subjective opinions. We found, however, that this is not necessarily a problem: Our models perform best on a *high-recall aggregate annotation* which includes all labels assigned by at least one annotator. Thus, we believe that the individual labels have value and are not, like generally assumed in crowdsourcing, noisy inputs suitable only as input for majority voting. In this vein, we publish all individual annotations. This enables further research on other methods of defining consensus annotations which may be more appropriate for specific downstream tasks. More generally, we will make all annotations, resources and model implementations publicly available.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, BC, Canada.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings*, pages 196–205. Springer.

Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva. 2016. USFD at semeval-2016 task 6: Any-target stance detection on twitter with autoencoders. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 389–393, San Diego, California.

Sarah Busch. 2011. Capturing mood about daily weather from twitter posts. http://www.dialogueearth.org/2011/09/29/capturing-mood-about-daily-weather-from-twitter-posts.

Sarah Busch. 2012. Tracking the mood about gas prices on twitter: A case study. http://www.dialogueearth.org/2012/01/25/tracking-the-mood-about-gas-prices-on-twitter-a-case-study.

Kent Cavender-Bares. 2011. Preparing to extract weather mood from tweets. http://www.dialogueearth.org/2011/03/03/preparing-to-extract-weather-mood-from-tweets.

Domenic V. Cicchetti and Alvan R. Feinstein. 1990. High agreement but low kappa: II. resolving the paradoxes. *Journal of clinical epidemiology*, 43:551–558.

Joana Costa, Catarina Silva, Mario Antunes, and Bernardete Ribeiro. 2014. Concept drift awareness in twitter streams. In *13th International Conference on Machine Learning and Applications*, pages 294–299.

Marcelo Dias and Karin Becker. 2016. INF-UFRGS-OPINION-MINING at SemEval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 378–383, San Diego, California.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 213–239, Palo Alto, California, USA.

Peter S. Dodds, Kameron D. Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12).

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1012–1017, Austin, Texas.

Paul Ekman. 1999. Basic emotions. In M Dalgleish, T; Power, editor, *Handbook of Cognition and Emotion*. John Wiley & Sons, Sussex, UK.

Lucie Flekova and Iryna Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2029–2041, Berlin, Germany.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *International Workshop on the Web and Databases (WebDB 2009)*, Providence, Rhode Island, USA.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

Yulan He Hassan Saif, Miriam Fernandez and Harith Alani. 2013. Evaluation Datasets for Twitter Sentiment Analysis: A survey and a new dataset, the STS-Gold. In *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013)*, pages 9–21, Turin, Italy.

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

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, Seattle, WA, USA.

Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *Proc. of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC)*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*, San Diego, CA, USA.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Roman Klinger. 2017. Does optical character recognition and caption generation improve emotion detection in microblog posts? In *Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings*, pages 313–319, Cham. Springer International Publishing.

Himabindu Lakkaraju, Chiranjib Bhattacharyya, Indrajit Bhattacharya, and Srujana Merugu. 2011. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 498–509, Mesa, Arizona, USA.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations*, Scottsdale, AZ, USA.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of WASSA at EMNLP*, Copenhagen, Denmark.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3).

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480 – 499.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California. Association for Computational Linguistics.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(July–August):344–350.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491, Avignon, France.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3806–3813, Istanbul, Turkey.

Cicero dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland.

Klaus Scherer and Harald Wallbott. 1997. The ISEAR questionnaire and codebook. Geneva Emotion Research Group.

David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2009. Tweet the debates: Understanding community annotation of uncollected sources. In *Proceedings of the First SIGMM Workshop on Social Media*, pages 3–10, Beijing, China.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63, Edinburgh, Scotland.

Damiano Spina, Edgar Meij, Maarten de Rijke, Andrei Oghina, Minh Thuong Bui, and Mathias Breuss. 2012. Identifying entity aspects in microblog posts. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1089–1090, New York, NY, USA. ACM.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, pages 121–136, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING*

2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3298–3307, Osaka, Japan.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science Technology*, 63(1):163–173.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, pages 587–592, Washington, DC, USA.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 Task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3).

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3485–3495, Osaka, Japan.