An Analysis of Annotated Corpora for Emotion Classification in Text

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

Abstract

Several datasets have been annotated and published for classification of emotions. They differ in several ways: (1) the use of different annotation schemata (*e. g.*, discrete label sets, including *joy*, *anger*, *fear*, or *sadness* or continuous values including *valence*, or *arousal*), (2) the domain, and, (3) the file formats. This leads to several research gaps: supervised models often only use a limited set of available resources. Additionally, no previous work has compared emotion corpora in a systematic manner. We aim at contributing to this situation with a survey of the datasets, and aggregate them in a common file format with a common annotation schema. Based on this aggregation, we perform the first cross-corpus classification experiments in the spirit of future research enabled by this paper, in order to gain insight and a better understanding of differences of models inferred from the data. This work also simplifies the choice of the most appropriate resources for developing a model for a novel domain. One result from our analysis is that a subset of corpora is better classified with models trained on a different corpus. For none of the corpora, training on all data altogether is better than using a subselection of the resources. Our unified corpus is available at http://www.ims.uni-stuttgart.de/data/unifyemotion.

Title and Abstract in German

Eine Analyse von annotierten Korpora zur Emotionsklassifizierung in Text

Es existieren bereits verschiedene Textkorpora, welche zur Erstellung von Modellen für die automatische Emotionsklassifikation erstellt wurden. Sie unterscheiden sich (1) in den unterschiedlichen Annotationsschemata (z.B. diskrete Klassen wie *Freude*, *Wut*, *Angst*, *Trauer* oder kontinuierliche Werte wie *Valenz* und *Aktivierung*), (2) in der Domäne, und, auf einer technischen Ebene, (3) in den Dateiformaten. Dies führt dazu, dass überwacht erstellte Modelle typischerweise nur einen Teil der verfügbaren Ressourcen nutzen sowie kein systematischer Vergleich der Korpora existiert. Hier setzt unsere Arbeit mit einem Überblick der verfügbaren Datensätze an, welche wir in ein gemeinsames Format mit einem einheitlichen Annotationsschema konvertieren. Darauf aufbauend führen wir erste Experimente durch, in dem wir auf Teilmengen der Korpora trainieren und auf anderen testen. Dies steht im Sinne zukünftiger, durch unsere Arbeit ermöglichten Analysen, die Unterschiede zwischen den Annotationen besser zu verstehen. Des Weiteren vereinfacht dies die Wahl einer angemessenen Ressource für die Erstellung von Modellen für eine neue Domäne. Wir zeigen unter anderem, dass die Vorhersagen für einige Korpora besser funktioniert, wenn ein Modell auf einer anderen Ressource trainiert wird. Weiterhin ist für kein Korpus die Vorhersage am besten, wenn alle Daten vereint werden. Unser aggregiertes Korpus ist verfügbar unter http://www.ims.uni-stuttgart.de/data/unifyemotion.

1 Introduction

Emotion detection and classification in text focuses on mapping words, sentences, and documents to a set of emotions following a psychological model such as those proposed by Ekman (1992), Plutchik (1980) or Russell (1980), *inter alia*. The task has emerged from a purely research oriented topic to playing a role in a variety of applications, which include dialog systems (chatbots, tutoring systems), intelligent agents, clinical diagnoses of mental disorders (Calvo et al., 2017), or social media mining.

As the variety of applications is large, the set of domains and differences in text is large. An early work, motivated by the goal to develop an empathic storyteller for children stories, is the corpus creation and modelling of emotions in tales by Alm et al. (2005). Afterwards, the idea has been transferred to the Web, namely blogs (Aman and Szpakowicz, 2007), and microblogs (Schuff et al., 2017; Mohammad, 2012; Wang et al., 2012). A different domain under consideration are news articles: Strapparava and Mihalcea (2007) focus on emotions in headlines. It can be doubted that emotions are expressed in a comparable way in these different domains: Journalists ideally tend to be objective when writing articles, authors of microblog posts need to focus on brevity, and one might assume that emotion expressions in tales are more subtle and implicit than, for instance, in blogs. Therefore, the transfer across emotion recognition models is, presumably, challenging. The most straight-forward alternative is, however, to build resources from scratch, a costly process. Given this situation, it remains unclear, given a novel domain for which an emotion recognition system should be developed, how to start. Next to the methodological issues we just discussed, another challenge is to acquire and compare available corpora.

With this paper, we aim at contributing to all these challenges. We support future research by comparing existing datasets, exploring how they complement each other, and mapping them to a common format such that future emotion detection can benefit from being developed on different domains and annotation schemata. In addition, we perform cross-corpus experiments by training classifiers on each dataset and evaluating them on others. One challenge here is that corpora are from different domains and are annotated following different guidelines and schemata. We aim at helping to make decisions which support the best model development for a future domain by selecting appropriate corpora. Parameters to be taken into account are the source of the text (*e. g.*, blogs, news, social media), the annotation schema (*e. g.*, Plutchik, Ekman, subsets of them), and the annotation procedure (*e. g.* crowdsourcing, self-reporting, expert-based, distant labeling). Our main contributions are therefore (1) to describe existing resources, (2) to evaluate which state-of-the-art classifiers perform well when trained on one dataset and applied on another, (3) to evaluate which datasets generalize best to other domains, and (4) to compare the datasets qualitatively and quantitatively. To achieve our goals and as additional support for future research, we unify all available datasets in a common file format. We provide a script that downloads and converts the datasets, and instructs on how to obtain datasets where the license requires no redistributions. Our resources are available via `http://www.ims.uni-stuttgart.de/data/unifyemotion`. We aim at keeping the unified corpus up to date in the future.

2 Background & Related work

In the following, we discuss differences in psychological models and annotation schemata (Section 2.1), annotation procedures (Section 2.2), different domains and topics (Section 2.3), and different prediction methods (Section 2.4). An overview on the resources and previous work is shown in Table 1. In addition to this, we recommend the surveys by Munezero et al. (2014) and Santos and Maia (2018).

2.1 Emotion Models in Psychology & Annotation Schemata

Emotion models are still debated in psychology (Barrett et al., 2018; Cowen and Keltner, 2018). We do not contribute to these debates but focus on the main theories in psychology and natural language processing (NLP): discrete and finite sets of emotions (categorical models) and combinations of different continuous dimensions (dimensional models).

Early work on emotion detection (Alm et al., 2005; Strapparava and Mihalcea, 2007) focused on conceptualizing emotions by following Ekman's model which assumes the following six basic emotions: *anger, disgust, fear, joy, sadness* and *surprise* (Ekman, 1992). Suttles and Ide (2013), Meo and Sulis

(2017), and Abdul-Mageed and Ungar (2017) follow the *Wheel of Emotion* (Plutchik, 1980; Plutchik, 2001) which also considers emotions as a discrete set of eight basic emotions in four opposing pairs: *joy–sadness*, *anger–fear*, *trust–disgust*, and *anticipation–surprise*, together with emotion mixtures.

Dimensional models were more recently adopted in NLP (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017a; Buechel and Hahn, 2017b): The circumplex model (Russell and Mehrabian, 1977) puts affective states into a vector space of valence (corresponding to sentiment/polarity), arousal (corresponding to a degree of calmness or excitement), and dominance (perceived degree of control over a given situation). Any emotion is a linear combination of these.

2.2 Annotation Procedures

A standard way to create annotated datasets is via *expert annotation* (Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2007; Ghazi et al., 2015; Li et al., 2017; Schuff et al., 2017; Li et al., 2017). However, having an expert annotate a statement means that they must estimate the private state of the author. Therefore, the creators of the ISEAR dataset follow a different, but similar, route making use of *self reporting*: subjects are asked to describe situations associated with a specific emotion (Scherer and Wallbott, 1994). This approach can be considered an annotation by experts in their own right.

Crowdsourcing, for instance using the platforms Amazon's Mechanical Turk[1] or CrowdFlower[2], is another way to acquire human judgments. Crowdsourcing often lacks sufficient quality control but some popular datasets have been successfully acquired with this approach, *e. g.*, the dataset released by Crowdflower for Cortana[3] or the datasets constructed by Milnea et al. (2015) and Lapitan et al. (2016). Another example is the dataset by Mohammad (2012), who design two detailed online questionnaires and annotate tweets by crowdsourcing.

Lastly, social networks play a central role in data acquisition with *distant supervision* (also called *self-labeling* in this context), because they provide a quick and cheap way to get large amounts of noisy data annotated by writers or readers (Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017; De Choudhury et al., 2012; Liu et al., 2017). For example, on Twitter one could add a "#joy" hashtag to a happy tweet or on Facebook one could tag personal posts with a "feeling" and people can show an emotional "surprised reaction". In this last example, two levels of annotation are provided that are relevant to emotion analysis, namely both the reader's and the writer's emotional state. Accessing this information is comparably straight-forward in these social network platforms. More challenging is to acquire such data for other domains. Buechel and Hahn (2017a) and Buechel and Hahn (2017b) look specifically into distinguishing between writers' and readers' emotion expressions.

It should be noted that some of these approaches exist in parallel to previous research in assessing emotion states of people, despite the fact that standardized psychological instruments exist (Bradley and Lang, 1994).

2.3 Domains and Topics

Previous work on emotion detection focuses on different domains and topics, *e. g.*, descriptions of *self reported emotional events* (Scherer and Wallbott, 1994), *news* (Lei et al., 2014; Buechel et al., 2016), *news headlines* (Strapparava and Mihalcea, 2007), *blogs* (Aman and Szpakowicz, 2007), *tales* (Alm et al., 2005), *micro-blog posts* (*i. e.*, tweets) (Wang et al., 2012) to different domains, such as *health*, *politics* (Mohammad, 2012), and *stock markets* (Bollen et al., 2011).

An early example and one of the first initiatives of emotion classification is the work by Aman and Szpakowicz (2007), who use *blog posts*, sampled without taking a specific topic into account. They identify the emotion, category, intensity and cue words and phrases. Mishne and de Rijke (2005), Balog et al. (2006) and Nguyen et al. (2014) works on LiveJournal[4] data to develop predictive models for moods.

Similarly, *user-generated data in social media* has been a subject of research. Mohammad et al. (2015) and Mohammad and Bravo-Marquez (2017b) annotate electoral *tweets* for sentiment, intensity, semantic

[1] https://www.mturk.com/

[2] https://www.crowdflower.com/

[3] https://www.crowdflower.com/data/sentiment-analysis-emotion-text/

[4] https://www.livejournal.com

roles, style, purpose and emotions. De Choudhury et al. (2012) identify more than 200 moods frequent on Twitter. Mohammad (2012), Mohammad et al. (2015), Wang et al. (2012), Volkova and Bachrach (2016) make use of Twitter distantly labeled data. Recently, Liu et al. (2017) analyzed the role of context that grounds sentiment in *tweets*, and looked into whether the effect of weather and news events relate to the emotion expressed in a given tweet. EmoNet is claimed to be the largest dataset constructed of *tweets* (Abdul-Mageed and Ungar, 2017) .

Twitter is often the preferred subject of research as it is easy to use and has a well-documented API. However, Facebook is also used, *e. g.*, Preoţiuc-Pietro et al. (2016) create a dataset of *Facebook posts* and train prediction models for valence and arousal. Pool and Nissim (2016) and Krebs et al. (2017) make use of the reaction feature in Facebook to collect labeled data for distant supervision of a classifier. A different approach within the same domain was used by Polignano et al. (2017) who labeled posts with emoticons mapped to Ekman's model.

Data in social media can be in the form of *dialogues*. Li et al. (2017) manually label a dataset of conversations. Wang et al. (2016) introduce EmotionPush, a system that automatically conveys the emotion of received text on mobile devices, deployed on Facebook's messenger app. From the same domain, but on a different topic is the study of patients' emotional states dynamics expressed by their Facebook posts (Lombardo et al., 2017).

Motivated by the work of literary scholars is the creation of datasets to study emotion in *literature*. One of the first datasets is the tales corpus by Alm et al. (2005). Kim et al. (2017) investigate the relationship between literary genres and emotions.

2.4 Methods used in Emotion Identification

Emotion classification is commonly phrased as text classification. As text classification in general, the array of methods seen for emotion classification can be divided into rule-based methods and machine learning, which we discuss in the following.

2.4.1 Rule-based Algorithms

Rule-based text classification typically builds on top of lexical resources of emotionally charged words. These dictionaries can originate from crowdsourcing or expert curation. Examples include WordNet-Affect (Strapparava et al., 2004) and SentiWordNet (Esuli and Sebastiani, 2007), both of which stem from expert annotation. Partly built on top of them is the NRC Word-Emotion Association Lexicon (Mohammad et al., 2013), which uses the eight basic emotions (Plutchik, 1980). Warriner et al. (2013) use crowdsourcing to assign values of valence, arousal, and dominance (Russell, 1980).

Another related category of lexical resources which has been used for emotion analysis is *concreteness* and *abstractness* (Köper et al., 2017). Brysbaert et al. (2014) publish a lexicon based on crowdsourcing, where the task was to assign a rating from 1 to 5 of the concreteness of 40,000 words. Similarly, Köper and Schulte im Walde (2016) automatically generate affective norms of abstractness, arousal, imageability, and valence for 350,000 lemmas in German. Lastly, the Linguistic Inquiry and Word Count (LIWC) is a set of 73 lexicons (Pennebaker et al., 2001), built to gather aspects of lexical meaning regarding psychological tasks. Dictionary and rule-based approaches are particularly common in the field of digital humanities due to their transparency and straightforward use.

2.4.2 Machine Learning

A performance improvement over dictionary lookup has been observed with supervised feature-based learning. Common features include word n-grams, character n-grams, word embeddings, affect lexicons, negation, punctuation, emoticons, or hashtags (Mohammad, 2012) . This feature representation is then usually used as input to feed classifiers such as naive Bayes, SVM (Mohammad, 2012), MaxEnt and others to predict the relevant emotion category (Aman and Szpakowicz, 2007; Alm et al., 2005). Similarly to the paradigm shift in sentiment analysis, from feature-based modelling to deep learning, state-of-the-art models for emotion classification are often based on neural networks (Barnes et al., 2017). Schuff et al. (2017) applied models from the classes of CNN, BiLSTM (Schuster and Paliwal, 1997), and LSTM (Hochreiter and Schmidhuber, 1997) and compare them to linear classifiers (SVM and MaxEnt), where the

Dataset	Granularity	Annotation	Size	Topic	Source	Avail.
AffectiveText	headlines	E + V	1,250	news	Strapparava (2007)	D-U
Blogs	sentences	E + ne + me	5,025	blogs	Aman (2007)	R
CrowdFlower	tweets	E + CF	40,000	general	Crowdflower (2016)	D-U
DailyDialogs	dialogues	E	13,118	multiple	Li et al. (2017)	D-RO
Electoral-Tweets	tweets	P	4,058	elections	Mohammad (2015)	D-RO
EmoBank	sentences	V+A+D	10,548	multiple	Buechel (2017a)	CC-by4
EmoInt	tweets	E − DS	7,097	general	Mohammad (2017b)	D-RO
Emotion-Stimulus	sentences	E + shame	2,414	general	Ghazi et al. (2015)	D-U
fb-valence-arousal	faceb. posts	V+A	2,895	questionnaire	Preoţiuc (2016)	D-U
Grounded-Emotions	tweets	HS	2,585	weather/events	Liu et al. (2017)	D-U
ISEAR	descriptions	E + SG	7,665	events	Scherer (1994)	GPLv3
Tales	sentences	E	15,302	fairytales	Alm et al. (2005)	GPLv3
SSEC	tweets	P	4,868	general	Schuff et al. (2017)	D-RO
TEC	tweets	E ±S	21,051	general	Mohammad (2012)	D-RO

Table 1: Selection of resources for emotions analysis. Ann. refers to the following annotation schemata: [E] *Ekman: anger, disgust, fear, joy, sadness, surprise*, [P] *Plutchik: anger, disgust, fear, joy, sadness, surprise, trust, anticipation*, [CF] *enthusiasm, fun, hate, neutral, love, boredom, relief, empty*, [DS] *disgust, surprise*, [JS] happy, sad, [V] *valence*, [A] *arousal*, [D] *dominance*, [SG] *shame, guilt*, [±S] *positive surprise, negative surprise*, [ne] *no emotion* [me] *mixed emotion* and Availability refers to the following [D-RO] *Available to download, research only*, [D-U] *Available to download, unknown licensing*, [R] *Available upon request*, [GPLv3] *GNU Public License version 3*, [CC-by 4] *Creative Commons Attribution version 4.0*

BiLSTM show best results with the most balanced precision and recall. Abdul-Mageed and Ungar (2017) claim the highest F_1 following Plutchik's emotion model with gated recurrent unit networks (Chung et al., 2015).

One approach to tackle sparsity of datasets is transfer learning; to make use of similar resources and then transfer the model to the actual task. A recent successful example for this procedure is Felbo et al. (2017) who present a neural network model trained on emoticons which is then transfered to different downstream tasks, namely the prediction of sentiment, sarcasm, and emotions.

3 Unified Dataset of Emotion Annotations

In this section, we describe each dataset we aggregate in our unified corpus. We provide a brief description and then show how the different schemata are merged. Please note that our interpretation might differ from the author's original description (though we aimed at avoiding that).

3.1 Datasets Overview

AffectiveText The dataset AffectiveText published by Strapparava and Mihalcea (2007) is built on news headlines and consists of 1,250 instances. The main goal of this resource is the classification of emotions and valence in news headlines. The annotation schema follows Ekman's basic emotions, complemented by valence. It is multi-label annotated via expert annotation and can be freely downloaded, the license is not specified. The emotion categories are assigned a score from 0 to 100. Training/developing data amounts to 250 annotated headlines, while systems are evaluated on another 1,000 instances.

Blogs This dataset, published by Aman and Szpakowicz (2007) consists of 5,205 sentences from 173 blogs. Instances are annotated with one emotion label each, emotion intensity and emotion indicators. The annotation schema for the emotion category corresponds to the six fundamental Ekman emotions to which *no emotion* is added. This resource can be obtained through contacting the authors.

CrowdFlower The dataset "The Emotion in Text, published by CrowdFlower" consists of 39,740 tweets. Part of this data has been used in Microsoft's Cortana Intelligence Gallery. The set of labels is non-standard (see Table 4). It is annotated via crowdsourcing with one label per tweet and can be freely downloaded, the license is not specified. The data is comparably noisy.

DailyDialogs The dataset, published by Li et al. (2017), is built on conversations and consists of 13,118 sentences. The annotation schema follows Ekman, complemented by "no emotion". It is single label annotated via expert annotation and can be freely downloaded for research purposes. This dataset has additional annotations for communication intention and topic.

Electoral-Tweets The dataset, published by Mohammad et al. (2015), targets the domain of elections. It consists of over 100,000 responses to two detailed online questionnaires (the questions targeted emotions, purpose, and style in electoral tweets). The tweets are annotated via crowdsourcing. The set of labels for emotion is non-standard (see Table 1). In addition to document-level annotations, tweets are annotated with emotion words. It can be freely downloaded for research purposes.

EmoBank The dataset published by Buechel and Hahn (2017a) builds on multiple genres and domains. It consists of 10,548 sentences where each sentence was manually annotated according to both the emotion which is expressed by the writer, and the emotion which is perceived by the readers. The annotations are according to the valence-arousal-dominance model. A subset of the corpus is AffectiveText, which makes this dataset a good resource to design models that map between both discrete or dimensional representations.

EmoInt The EmoInt published by Mohammad and Bravo-Marquez (2017a) builds on social media content that amounts to 7,097 tweets altogether. The main goal of this resource is to associate text with various intensities of emotion. The tweets are annotated via crowdsourcing with intensities of *anger*, *joy*, *sadness*, and *fear*, while most tweets are only annotated with one emotion. It can be freely downloaded for research purposes.

Emotion-Stimulus The Emotion-Stimulus dataset published by Ghazi et al. (2015) consists of 820 sentences which are annotated both with emotions and their causes, and 1,549 sentences which are marked only with their emotion. The set of labels used for annotation consists of Ekman's basic emotions to which *shame* is added. The main goal of this resource is to predict the cause of an emotion in the text. It is annotated using FrameNet's emotions-directed frame with one emotion label per sentence. It is available for download for research purposes.

fb-valence-arousal The fb-valence-arousal dataset published by Preoţiuc-Pietro et al. (2016) is built on Facebook posts. It consists of 2,895 posts stratified by age and gender. The main goal of this resource is to train prediction models for *valence* and *arousal*. Each message is written by a distinct user and all messages are from the same time interval. The posts are annotated with valence and arousal on a nine point scale via expert annotation. It is available for download.

Grounded-Emotions The dataset published by Liu et al. (2017) is built on social media and consists of 2,557 single labeled instances published by 1,369 unique users. The main goal of this resource is to put emotions into context of other factors including weather, news events, social network, user predisposition, and timing. The set of labels is *happy* and *sad*. The tweets are annotated by the authors. The information is used in experiments aiming at showing the role played by external context in predicting emotions.

ISEAR The "International Survey on Emotion Antecedents and Reactions" dataset published by Scherer and Wallbott (1994) is built by collecting questionnaires answered by people with different cultural backgrounds. These people report on their own emotional events. The final dataset contains reports by approximately 3,000 respondents, for a total of 7,665 sentences labeled with single emotions. The labels are *joy, fear, anger, sadness, disgust, shame*, and *guilt*. It is available for download.

SSEC The Stance Sentiment Emotion Corpus published by Schuff et al. (2017) is an annotation of the SemEval 2016 Twitter stance and sentiment dataset (Mohammad et al., 2017). It consists of 4,868 tweets. The main goal of this resource is to enable further research on the relations between emotions and other factors. It is annotated via expert annotation with multiple emotion labels per tweet following Plutchik's fundamental emotions. An additional feature of this resource is that they not only provide a majority annotation but publish the individual information for all annotators.

Tales The Tales corpus published by Alm et al. (2005) is built on literature and consists of 15,302 sentences from 185 fairytales by B. Potter, H.C. Andersen and the brothers Grimm. Out of these 15,302 sentence, all annotators agree only on 1,280. The main goal of this resource is to build

	joy		anger		sadness		disgust		fear		surprise	
Dataset	o	m	o	m	o	m	o	m	o	m	o	m
AffectiveText	619	619	374	374	650	650	253	253	537	537	787	787
Blogs	848	536	884	179	883	173	882	172	861	115	847	115
CrowdFlower	5,209	9,220	110	1,421	5,165	5,123	—	179	8,459	8,430	2,187	2,177
DailyDialogs	12,885	12,885	1,022	1,022	1,150	1,150	353	353	74	174	1,823	1,823
Electoral-Tweets	267	349	300	569	34	31	1,937	1,638	80	91	259	251
EmoInt	1,616	1,616	1,701	1,701	1,533	1,533	—	—	2,252	2,252	—	–
Emotion-Stimulus	479	479	483	483	575	575	95	95	423	423	213	213
Grounded-Emotions	1,530	1,530	—	—	1,055	1,055	—	—	—	—	—	—
ISEAR	1,094	1,094	1,096	1,096	3,285	3,285	1,096	1,096	1,095	1,095		—
SSEC	2,067	2,067	2,902	2,902	2,644	2,644	2,183	2,183	1,840	1,840	1,108	1,108
Tales	107	107	195	195	116	116	14	14	111	111	90	90
TEC	8,240	8,237	1,555	1,555	3,830	3,830	761	761	2,816	2,816	3,849	3,849
Total	34,961	38,739	10,622	11,497	21,920	20,165	7,574	6,744	16,708	17,884	11,162	10,413

Table 2: The distribution of categories across the datasets, limited to *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. Non-availability of a class in a set is marked with —. [o]: original distributions, without taking high agreement annotations; [m]: counts after mapping to our labels (see Table 4).

emotion classifiers for literature. The annotation schema consists of Ekman's six basic emotions. In the final data the labels *angry* and *disgust* are merged. It can be freely downloaded for research purposes.

TEC The Twitter Emotion Corpus published by Mohammad (2012) is built on social media. It consists of 21,051 tweets. The main goal of this resource is answering the question if emotion-word hashtags can successfully be used as emotion labels. The annotation schema corresponds to Ekman's model of basic emotions. They collected tweets with hashtags corresponding to the six Ekman emotions: *#anger*, *#disgust*, *#fear*, *#happy*, *#sadness*, and *#surprise*, therefore it is distantly single-label annotated. It can be freely downloaded for research purposes.

3.2 Analysis

The majority of resources consists of user generated data. The biggest dataset is CrowdFlower with 39,740 annotated tweets, followed by TEC with 21,051 tweets, and Blogs with 15,000 annotated sentences from blog posts. Out of all 14 resources, 9 are annotated with the six fundamental emotions defined by Ekman, with small variations. SSEC and Electoral-Tweets follow Plutchik's model. Electoral-Tweets is annotated with 19 emotions and the authors provide a mapping to Plutchik's set (which we follow in the aggregation). Non-fundamental emotions are annotated in CrowdFlower (*fun, worry, enthusiasm, and love*).

Not only the size, also the distribution of labels is different in the corpora. Table 2 shows the distribution for Ekman's emotions before and after having applied the mapping to a unique set of emotion labels (see Table 4 in the Appendix A). In many corpora, *joy* is the dominating emotion, followed by *sadness*, *surprise*, and *anger*. Exceptions are SSEC, Electoral-Tweets, and EmoInt, in which negative emotions are more frequent. In SSEC, this is because of its origin as a stance dataset. Similarly, Electoral-Tweets shows a polarizing nature of political debates with *disgust* and *anger* being more common.

Figure 1 shows a quantitative similarity comparison of the data. We represent each dataset by its term distribution, taking the top 5,000 most common words from each dataset and calculating the cosine similarity across corpora (inspired by Ruder and Plank (2017) and Plank and Van Noord (2011)). Twitter corpora are more similar to each other (EmoInt and CrowdFlower are the most similar to TEC) than to other domains with the exception of SSEC, which is the most dissimilar to the other tweet datasets. DailyDialogs is more similar to the tweets than to ISEAR and Blogs.

The column *All* stands for the union of all datasets except the one that is being compared to. In this context, the most dissimilar towards the respective aggregated set is AffectiveText. The reason is that this is a small dataset compared to the tweet-based corpora and that it covers a specific topic, *headlines*. Grounded-Emotions is also notably dissimilar. Most similar to *All* is EmoInt, followed closely by TEC and Blogs, which covers blog posts and not tweets.

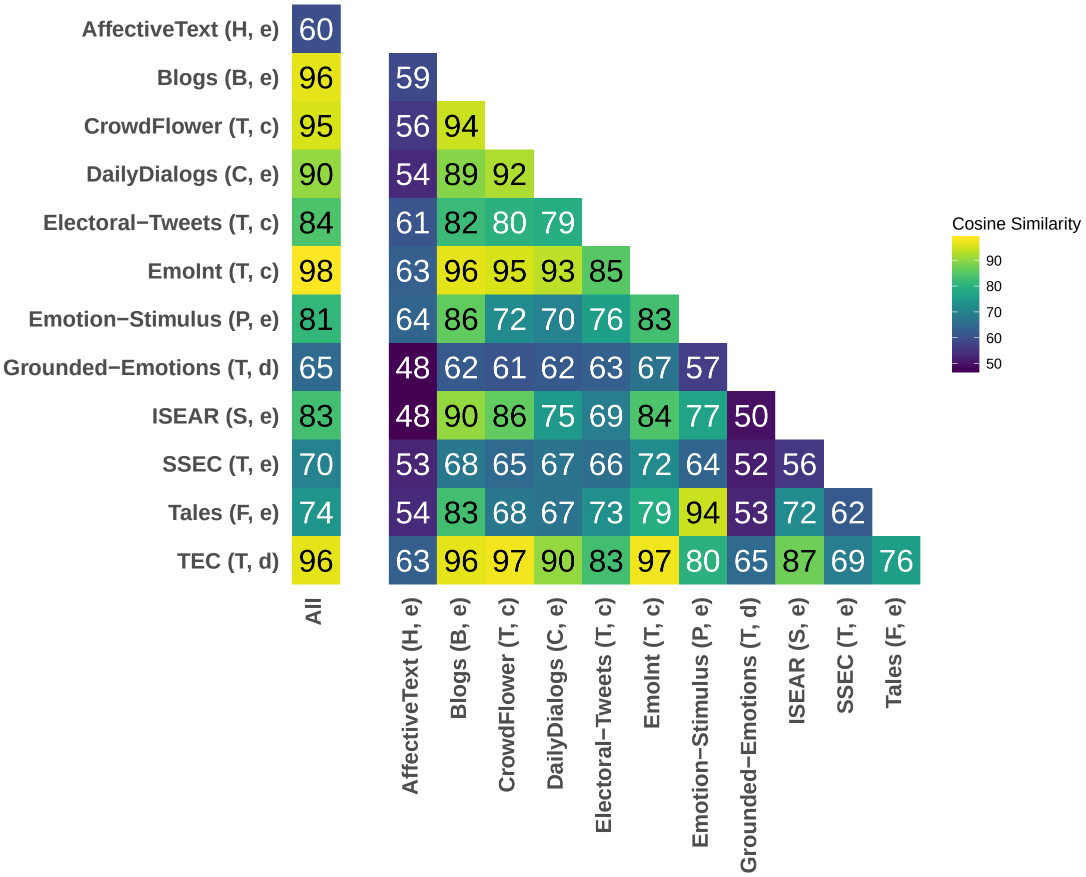

Figure 1: Similarity of corpora with cosine measure. T: tweets, F: tales, P: paragraphs, S: descriptions, H: headlines, B: blogpost, C: conversations; e: expert annotation, d: distant supervision, c: crowdsourcing.

3.3 Aggregation

To provide a standardized access to the datasets, we define *joy, anger, sadness, disgust, fear, trust, surprise, love, confusion, anticipation* and *noemo* as our common label set. The original resources additionally include other 44 labels that come from Electoral-Tweets and CrowdFlower. Where availble from the original publication, we follow proposed mappings (*e. g.*, Electoral-Tweets with 19 emotions and a mapping to Plutchik's model). Table 4 in Appendix A summarizes the mapping. We include *valence, arousal*, and *dominance* where annotated, however, we currently do not map the categorical emotion models onto the dimensional ones.

Each instance in the unified dataset contains, in addition, a unique id, the source corpus name, the text, and an assignment of a real number to each of the 11 emotion variables. In most datasets, each instance is only annotated discretely with single labels (exceptions are SSEC and AffectiveText). Therefore, most instances have exactly one label marked with a 1.0. For the multi-labeled datasets, several emotions can be marked. For datasets with annotated emotion intensity, values can range between 0 and 1. For datasets with multiple annotator information, we follow the recommendations by the original authors. For SSEC, that means, accepting a label if at least one annotator assigned it. For Tales, the authors provide a gold annotation, in which *angry* and *disgust* are merged. We handle them separately, and accept a label if and only if all annotators agree. For Blogs, we take the examples of the dataset with the high agreement annotations.

Next to these fundamental attributes, we provide the domain and annotation style information, as well as additional, dataset specific attributes (*e. g.*, the story from which an instance originates, in the Tales corpus). Our unified data includes all datasets for which the licenses are available; as some datasets are not redistributable, but free to download, we provide a script that interactively downloads and converts the existing resources into our unified format. An excerpt of the data is depicted in Appendix B.

4 Experiments

We perform experiments in the following settings: Firstly, we perform a within-corpus emotion classification (training on one corpus and testing on the same, using cross-validation[5]). Secondly, we do pairwise-corpus evaluation: training on the entire data of one corpus and evaluating on all the data of a different one, for all corpus pairs. This includes the use of the aggregated corpus, but for this experiment excluding the test corpus – this corresponds to a cross validation setting in which the subsets correspond to corpora.

4.1 Experimental settings

Previous methods have shown that linear classifiers are nearly *en par* with neural methods (Schuff et al., 2017). We therefore use maximum entropy classifiers as implemented in scikit-learn (Pedregosa et al., 2011) with bag-of-words (BOW) features for these experiments for simplicity and easy reproducibility. We use L2 regularization and balance the classes in the training data with instance weights. Training/test splits are 80%/20%. Cross-validation is 10-folds stratified.

For datasets in which the labels do not align following our mapping, we use the intersection of labels in the train and test data. We do not discard any instances. For datasets that are designed for other tasks than emotion classification such as EmoInt and Emotion-Stimulus, we do not change the setting of our classification task.

For experiments in which we move from a *multi-label setting* (more than one emotion can hold per instance) to a *single-label setting* (only one emotion holds), we train multiple binary classifiers from which we only accept, in the prediction phase, the highest scoring emotion. For experiments in which we move from a *single-label setting* to a *multi-label setting*, we as well train separate binary classifiers and accept all emotions for which the binary classifiers output one. The case *multi-label* to *multi-label* works analogously. For *single-label* to *single-label*, we use one multi-class MaxEnt model.

4.2 Results

Within-corpus emotion classification. The results for our first experiment are shown per emotion in Table 3 (where we restrict the results to the six fundamental emotions defined by Ekman) and in the diagonal of Figure 2. These should be interpreted in context to the similarity analysis in Figure 1. We see that some datasets and domains are more difficult to be modeled than others. The "easiest" dataset seems to be Emotion-Stimulus, followed by EmoInt. The reason for the high scores lies in the fact that both datasets are constructed for different tasks (stimulus and intensity prediction). As such, our task does not suit these two very well.

Next datasets with comparably high performance measures are Blogs and DailyDialogs. In contrast, CrowdFlower and Electoral-Tweets seem to be the most challenging in the within-corpus setting. For CrowdFlower, the results are due to the larger label set, which makes the task more difficult. Mostly, the emotions that occur less frequently (like surprise) show lower results than the ones occurring frequently (like *joy* and *sadness*). In addition, manual inspection shows that this data is comparably noisy. This is also backed by the general observation that our model performs in general worse on Twitter data than on most other domains.

In terms of annotation procedures, these experiments allow almost for no judgement, since most of the datasets use expert annotation and we only have few examples for the other two ways of annotation (crowdsourcing and distant supervision) being used. However, we could observe that the crowdsourced datasets are more difficult which might be due to a more noisy annotation.

Cross-corpus emotion classification The in-corpus results (the diagonal in Figure 2) shows higher F_1-scores than the cross-corpus results. The exception is Electoral-Tweets, where the same performance is observed by training on a different corpus, Blogs. Models trained on Twitter data perform slightly

[5] Some datasets came with designated train and test parts, but in order to treat all datasets equal, we chose to ignore that. However, in the aggregated dataset, the information about which part of the corpus an example is from is preserved in a special field.

	Joy			Sadness			Fear			Anger			Disgust			Surprise		
Dataset	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
AffectiveText	68	64	66	72	64	68	85	55	67	37	76	50	44	55	49	69	74	71
Blogs	63	70	67	42	32	36	67	43	53	61	31	41	73	51	60	24	28	25
CrowdFlower	42	35	38	26	28	27	32	30	31	21	29	24	6	23	9	9	9	9
DailyDialogs	43	63	51	13	47	20	10	33	16	12	44	18	8	37	13	16	56	24
Electoral-Tweets	37	23	28	23	35	27	26	42	32	36	36	36	52	35	42	15	25	18
EmoInt	94	92	93	83	81	82	85	90	88	90	86	88	0	0	0	0	0	0
Emotion-Stimulus	98	99	98	95	100	97	99	98	98	99	98	98	100	90	95	97	97	97
Grounded-Emotions	61	53	57	41	48	44	0	0	0	0	0	0	0	0	0	0	0	0
ISEAR	61	74	67	73	66	69	69	69	69	45	48	47	58	61	59	0	0	0
SSEC	67	67	67	59	89	71	45	72	55	79	74	76	64	62	63	44	55	49
Tales	34	35	34	24	30	26	27	28	27	28	34	31	16	21	18	19	28	23
TEC	67	69	68	46	44	45	58	56	57	41	40	40	27	26	26	54	50	52

Table 3: Results obtained via 10-fold crossvalidation in precision, recall, and F_1 micro-averaged of the MaxEnt classifier with BOW as features, reported per Ekman's fundamental emotion. Zeros denote that the respective class is not annotated in the respective data set. Note that a subset of datasets has more classes annotated than the provided Ekman's emotions.

better on other Twitter sets, with an exception of Electoral-Tweets, for which the distribution of labels is different, with *disgust* dominating the set.

It is notable that EmoInt, Emotion-Stimulus, Grounded-Emotions ISEAR, and SSEC are easier to classify (high performance when used for testing) while DailyDialogs, Blogs, CrowdFlower, and Tales are more informative: training on them and classifying other datasets leads to better results. Models trained on ISEAR and SSEC perform comparably well. DailyDialogs is classified best not by a classifier trained on itself, but by a classifier trained on Blogs.

We cannot recommend to train on Emotion-Stimulus and Grounded-Emotions as long as the specific properties of these datasets are not required. The models estimated on these data do not perform well on other sets. Note that this is not a quality judgement, Grounded-Emotions has different labels and Emotion-Stimulus was designed for a different purpose.

Note that the similarity measure is an approximation of mediocre quality for model performance, with a Pearson correlation of $r = 0.32$.

All vs. one cross-corpus emotion classification. The results for this setting can be seen in the *All* column of Figure 2. These results show which datasets are easier to classify, namely DailyDialogs, Blogs, and EmoInt. It might seem intuitive that adding more and diverse training data could be helpful in classifying almost all datasets. However, we can see in the results that this is not the case. Especially the multi-labeled datasets AffectiveText and SSEC together with the datasets that were the most transformed (*i. e.*, many labels unified) during the aggregation process, such as Electoral-Tweets and CrowdFlower are more difficult to classify while training on all the other datasets.

5 Conclusion & Future Work

Datasets annotated for emotion classification are important in emotion analysis research as they are used in many downstream tasks as well; having these datasets all in place reduces drastically the amount of work needed in preprocessing and transferring the needed data. Yet, it is a diverse collection of datasets driven by different psychological emotion models, on different domains, and approaches used in annotation. With this paper, we present the first survey on emotion datasets on text.

In addition to this literature review, we provide a unified, aggregated corpus to support future research on standardized data. The existence of such a benchmark opens up the possibility of other experiments, such as transfer learning and domain adaptation work, with focus on different domains and on different label sets. From the collected unified datasets one could learn how to select the most suitable dataset for a given new domain and evaluate it across different classification models, domains, and annotation procedures, easier than it was possible until now. Also having this open will help the emotion detection

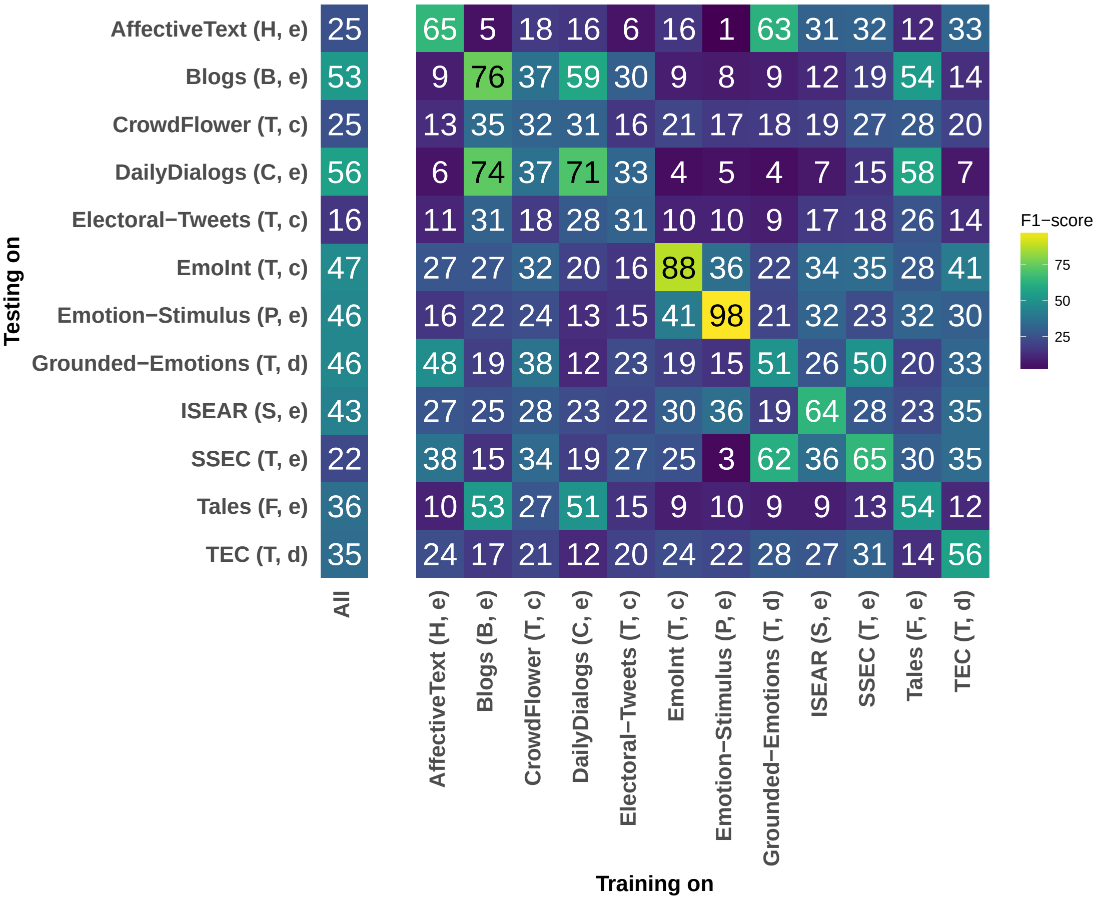

Figure 2: Results of MaxEnt in F_1 measure (micro-averaged) for all cross-corpus experiments. T: tweets, C: conversations, F: tales, P: paragraphs, S: descriptions, H: headlines, B: blogsposts; e: expert annotation, d: distant supervision, c: crowdsourcing.

task to become a standard task on which state-of-the-art methods used in general classification are tested upon, similarly to other tasks like sentiment classification, which plays this role already.

In addition, this work can be used by anyone who wants to explore the current state of the emotion analysis field. As future work we plan to release another version of the dataset in which the conversion between the different emotion models are added and to perform transfer learning experiments between datasets, domains, and annotation procedures. Furthermore, we propose to use the resource to qualitatively analyze the different realizations of emotions across annotation schemata and domains.

Acknowledgements

This research has been funded by the German Research Council (DFG), project SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1-1). We thank Evgeny Kim and Jeremy Barnes for fruitful discussions.

References

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)*, pages 718–728, Vancouver, Canada, July. Association for Computational Linguistics.

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, British Columbia, Canada, October. Association for Computational Linguistics.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Krisztian Balog, Gilad Mishne, and Maarten De Rijke. 2006. Why are they excited?: identifying and explaining spikes in blog mood levels. In *11th Conference of the European Chapter of the Association for Computational Linguistics: Demostrations*, pages 207–210. Association for Computational Linguistics.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Copenhagen, Denmark. Workshop at Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Lisa Feldman Barrett, Zulqarnain Khan, Jennifer Dy, and Dana Brooks. 2018. Nature of emotion categories: Comment on cowen and keltner. *Trends in Cognitive Sciences*, 22(2):97 – 99.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Sven Buechel and Udo Hahn. 2017a. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017b. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain, April. Association for Computational Linguistics.

Sven Buechel, Udo Hahn, Jan Goldenstein, Sebastian G. M. Händschke, and Peter Walgenbach. 2016. Do enterprises have emotions? In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 147–153, San Diego, California, June. Association for Computational Linguistics.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2067–2075, Lille, France, 07–09 Jul. PMLR.

Alan S. Cowen and Dacher Keltner. 2018. Clarifying the conceptualization, dimensionality, and structure of emotion: Response to barrett and colleagues. *Trends in Cognitive Sciences*, 22(4):274–276.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! exploring human emotional states in social media. In *Sixth international AAAI conference on weblogs and social media*, pages 66–73.

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

Andrea Esuli and Fabrizio Sebastiani. 2007. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, pages 1–26.

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, September. Association for Computational Linguistics.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 152–165, Cham. Springer International Publishing.

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

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada, August. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Copenhagen, Denmark. Workshop at Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Florian Krebs, Bruno Lubascher, Tobias Moers, Pieter Schaap, and Gerasimos Spanakis. 2017. Social emotion mining techniques for facebook posts reaction prediction. *arXiv preprint arXiv:1712.03249*.

Fermin Roberto Lapitan, Riza Theresa Batista-Navarro, and Eliezer Albacea. 2016. Crowdsourcing-based annotation of emotions in filipino and english tweets. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 74–82, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems*, 37:438–448.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Vicki Liu, Carmen Banea, and Rada Mihalcea. 2017. Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483, San Antonio, Texas, Oct.

Gianfranco Lombardo, Alberto Ferrari, Paolo Fornacciari, Monica Mordonini, Laura Sani, and Michele Tomaiuolo. 2017. Dynamics of emotions and relations in a facebook group of patients with hidradenitis suppurativa. In *International Conference on Smart Objects and Technologies for Social Good*, pages 269–278. Springer.

Rosa Meo and Emilio Sulis. 2017. Processing Affect in Social Media: A Comparison of Methods to Distinguish Emotions in Tweets. *ACM Transactions on Internet Technology*, 17(1):7:1–7:25.

David Milnea, Cecile Parisb, Helen Christensenc, Philip Batterhamc, and Bridianne ODeac. 2015. We feel: Taking the emotional pulse of the world. In *Proceedings 19th Triennial Congress of the IEA*, volume 9.

Gilad Mishne and Maarten de Rijke. 2005. Boosting web retrieval through query operations. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval*, pages 502–516, Berlin, Heidelberg. Springer Berlin Heidelberg.

Saif Mohammad and Felipe Bravo-Marquez. 2017a. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada, August. 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, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3):26:1–26:23, June.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: 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 (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Myriam D 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.

Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. 2014. Mood sensing from social media texts and its applications. *Knowledge and Information Systems*, 39(3):667–702, June. 00011.

W. Gerrod Parrott. 2000. *Emotions in Social Psychology*. Key Readings in Social Psychology. Psychology Press.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Barbara Plank and Gertjan Van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1566–1576. Association for Computational Linguistics.

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

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Marco Polignano, Marco de Gemmis, Fedelucio Narducci, and Giovanni Semeraro. 2017. Do you feel blue? detection of negative feeling from social media. In Floriana Esposito, Roberto Basili, Stefano Ferilli, and Francesca A. Lisi, editors, *AI*IA 2017 Advances in Artificial Intelligence*, pages 321–333, Cham. Springer International Publishing.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39. The COLING 2016 Organizing Committee.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15. Association for Computational Linguistics.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark, September. Association for Computational Linguistics.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Diana Santos and Belinda Maia. 2018. Language, emotion, and the emotions: A computational introduction. *Language and Linguistics Compass*, 12(6):e12279.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Copenhagen, Denmark. Workshop at Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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, June. Association for Computational Linguistics.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1083–1086.

Jared Suttles and Nancy Ide. 2013. Distant Supervision for Emotion Classification with Discrete Binary Values. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, number 7817 in Lecture Notes in Computer Science, pages 121–136. Springer Berlin Heidelberg, March. 00029.

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 (Volume 1: Long Papers)*, volume 1, pages 1567–1578.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter ”big data” for automatic emotion identification. In *SocialCom/PASSAT*, pages 587–592. IEEE.

Shih-Ming Wang, Chun-Hui Scott Lee, Yu-Chun Lo, Ting-Hao Huang, and Lun-Wei Ku. 2016. Sensing emotions in text messages: An application and deployment study of emotionpush. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 141–145, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

A Mapping of Labels to Unified Labels

Unified Label	Original Labels
anger	anger, angry, annoyance, fury, hostility, ag, hate
anticipation	anticipation, vigilence, interest, expectancy
confusion	confusion, indecision
disgust	disgust, dg, dislike, boredom, hate, disappointment, indifference
fear	fear, panic, terror, apprehension, fr, worry
joy	joy, happy, happiness, joyful, elation, hp, fun, enthusiasm, relief, serenity, calmness
love	love
noemo	noemo, neutral, ne, BLANK
sadness	sadness, sad, gloominess, grief, sorrow, sd, shame, guilt
surprise	surprise, uncertainty, amazement , su+, su−
trust	trust, acceptance, admiration, like

Table 4: Mapping of original labels to labels in unified dataset. We roughly follow the models by Plutchik (2001) and Parrott (2000)

B Example Excerpt from the Aggregated Corpus

```
{
  "id": 210599,
  "VAD": {
    "valence": null,
    "arousal": null,
    "dominance": null
  },
  "source": "ssec",
  "text": "He who exalts himself shall be humbled; and he who humbles himself shall
      be exalted. Matt 23:12. #SemST"
  "emotions": {
    "joy": 1,
    "anger": 1,
    "sadness": 0,
    "disgust": 1,
    "fear": 0,
    "trust": 1,
    "surprise": 0,
    "love": null,
    "noemo": null,
    "confusion": null,
    "anticipation": 0
     },
  "original_split": "test",
  "emotion_model": "Plutchik"
  "domain": "social-media/tweets",
  "labeled": "multilabeled"
}
```

Figure 3: Example excerpt from our aggregated and unified corpus.