Sentence and Clause Level Emotion Annotation, Detection, and Classification in a Multi-Genre Corpus

Shabnam Tafreshi, Mona Diab
George Washington University
Department of Computer Science, Washington D.C.
{shabnamt, mtdiab}@gwu.edu

Abstract

Predicting emotion categories (e.g. anger, joy, sadness) expressed by a sentence is challenging due to inherent multi-label smaller pieces such as phrases and clauses. To date, emotion has been studied in single genre, while models of human behaviors or situational awareness in the event of disasters require emotion modeling in multi-genres. In this paper, we expand and unify existing annotated data in different genres (emotional blog post, news title, and movie reviews) using an inventory of 8 emotions from Plutchik's Wheel of Emotions tags. We develop systems for automatically detecting and classifying emotions in text, in different textual genres and granularity levels, namely, sentence and clause levels in a supervised setting. We explore the effectiveness of clause annotation in sentence-level emotion detection and classification (EDC). To our knowledge, our EDC system is the first to target the clause level; further we provide emotion annotation for movie reviews dataset for the first time.

Keywords: emotion annotation, emotion analysis, multi-genre corpus

1. Introduction

Prediction of sentence-level emotion classification encompass a variety of applications such as modeling of human behaviors (Dodds and Danforth, 2010) and situational awareness in the event of disasters (Vo and Collier, 2013). As a precursor to our system development, we realize the diversity and non uniformity of existing resources with emotion tags, hence, we re-annotate existing resources in a unified framework, thereby covering multiple genres of text. The genres are as follows: emotional blog post (BLG), news headlines dataset (HLN), and movie review dataset (MOV). We present an approach and system that performs emotion detection and classification (EDC) on multiple levels of granularity, namely, sentence and clause levels. We expand the annotation scheme to cover both sentence and clause level annotations, as well as expand the emotion tag inventory from the typical Ekman 6 (Ekman, 1992) emotion labels (EK6) to 8 emotion labels based on Plutchik's Wheel of Emotions (Plutchik, 1962) (PL8).

In this study, we focus on the impact of clause-level annotation on the EDC task, which can be used effectively in a single-genre or multi-genre textual setting without significant performance loss. Similar to previous studies, we cast the EDC problem in a supervised setting. Evaluation of EDC in 10% held out data outperformed the baseline and gives the average accuracy of 81.1% and 71.3% for sentence and clause level respectively. EDC achieved better results compared to previous annotation of HLN and BLG datasets with EK6 emotion labels (average accuracy 54.7% and 73.8%). Accordingly, our contributions are as follows:

- A new set of annotation guidelines for emotion detection based on Plutchik's Wheel of Emotions.
- A uniformly annotated multi-genre data set (including old and new data) on two levels of granularity: sentence and clause levels.
- Two EDC systems on the sentence and clause levels for multiple genres leveraging clause-level annotation on sentence-level EDC systems.

The rest of this article is structured as follows: section 2. describes related work to the study; in section 3. we give data references, collection, annotation process and evaluation, and annotation challenges; section 4. explains the experiment setup and EDC description; and section 5. concludes and describes future direction of our study.

2. Related Work

Emotion detection in NLP has been studied on document, sentence, and phrase levels. Several studies investigated the problem in various data genres. We present studies most relevant to this paper. Aman and Szpakowicz (2007) collected and labeled BLG corpus using EK6 tags in sentence and phrase-level. Strapparava and Rada (2007), collected HLN set and labeled it using EK6 tags and valence, which, valence measures the polarity of each data point. HLN is used in SemEval 2007, task 14. Pang and Lee (2005) crawled web to collect MOV dataset to address rating inference problem. Mishne (2005) collected a set of blog posts - online diary entries - which include an indication of the writers's mood. Yan (2014) expanded the range of automatic emotion detection in microblogging text using three sampling strategies: random sampling, topics and events sampling, and sampling based on users. Abdul-Mageed and Lyle (2017) collected a large set of tweets using hashtags, they used Plutchik's Wheel of Emotions to create relevant hashtags, and the set is annotated using distant supervision method. To date, sentence-level emotion classification has been studied by a large group of researchers (Aman and Szpakowicz, 2007; Strapparava and Rada, 2007; Mishne, 2005; Yan, 2014; Das and Bandyopadhyay, 2010; Ghazi et al., 2010; Kim et al., 2010; Mohammad, 2012; Özbal and Pighin, 2013; Abdul-Mageed and Lyle, 2017), who ad-

dressed the EDC task on the document and sentence levels, to our knowledge, nobody investigated automatic tagging on the clause level and the impact of clause-level on sentence-level emotion classification, and that distinguish our work from previous works.

3. Data Description

We aim to create a multigenre corpus annotated with emotion tags on the clause and sentence level. We would like to cater to fine grained emotion detection with the goal of eventually building systems that detect emotion intensity. Toward that goal, we create a unified multigenre data set annotated on the clause and sentence levels. Moreover, we compared the typical EK6 to other tag sets that are more fine grained and well established in the psychology literature. We opted for Plutchik's Wheel of Emotions. Below is a detailed description of the data and the annotation process.

3.1. Corpus

We combined and annotated several previously annotated data sets on the sentence level for various types of emotions. The first data set is a emotional blog post (BLG) (Aman and Szpakowicz, 2007) where people typically express their emotions and opinions about social/personal events, politics, products, etc. This dataset comprises 4115 sentences. The second data set, a news headlines dataset (HLN) (Strapparava and Rada, 2007) crafted by creative people to possibly provoke emotions comprises 1250 sentences. Both BLG and HLN were annotated originally using the EK6 tag set. Finally, the third data set, a movie review dataset (MOV) (Pang and Lee, 2005) where people express their opinion about movies, sound tracks, and casts. The MOV data set contains 11,855 sentences. The MOV data set is annotated for sentiment intensity. The total number of sentences in the collection is 17,220. We extract clauses from the sentences in the three corpora using the Stanford parser (Klein and Manning, 2003) from the CoreNLP toolkit (Manning et al., 2014). In each sentence parse tree, we extract the labels, SBAR, SBARQ, etc. according to the Penn Treebank's clause labels of the parse trees (Marcus et al., 1993) identifys the sentence clauses. The total number of clauses corresponding to 17,220 sentences is 29,938. 7,458 of the sentences comprise a single clause. We refer to this sentence-level corpus as SBHM and clause-level corpus as CBHM.

3.2. Annotation Process

Annotating emotional data is a challenging task, since people perceive various experiences differently. This is expected to be the case especially when the data is extracted from social media platforms like forums and blogs. To develop appropriate emotion categories, we carried out our annotation procedure in two stages: a pilot stage and an annotation stage.

Pilot Stage: our work was guided by the following research questions:
(1) what emotion categories can be best suited for different genres in our corpus, what is the appropriate tag set for our multigenre corpus: Ekman's six basic emotions (EK6) or Plutchik's eight basic emotions (PL8)?
(2) In case of clause level annotation, what is the appropriate presentation method to the annotators?
To answer question (1), we set up an online survey. We selected 518 single clause sentences from the BHM corpus such that they equally represented the three underlying corpora BLG, MOV, HLN. Three annotators, graduate students, worked on the pilot data. We provided annotators with detailed guidelines regarding the task. We ran two pilot annotations: one asking annotators to use the EK6 tagset and the second where they were asked to use the PL8 tagset. Cases of disagreement between the annotators were discussed until a Fleiss Kappa K= 0.7 was reached for both pilot annotation exercises. The output of the pilot stage was an agreement to use the PL8 basic emotions, since it was a better reflection of the data. In addition, the annotators suggested adding the labels *interest, disappointment, confusion*, and *frustration*, but since these were not very frequently assigned (less than 2%), we decided to use the label *other-emotion* instead of adding these extra ones. We also added *no-emotion* to the tag set as an option available to annotators. Accordingly, based on feedback, we ended up with 10 labels including: PL8 set *joy, trust, anticipation, surprise, fear, sadness, disgust, anger*, *no-emotion*, and *other-emotion*. These annotations were collected on the sentence level. To address the second question, we further randomly selected 20 clauses testing how to demonstrate the clauses to the annotators. Based on a survey completed by 10 people, majority voted for marking clauses within each sentence and asking for an emotion tag, as opposed to showing the clauses in isolation without context. Hence, when annotating clauses, we mark each clause within its sentence, and provide it to the annotator. Below we demonstrate an example, clauses are marked as underline text:
Clause-1: It takes a really long , slow and dreary time to dope out what TUCK EVERLASTING is about .
Clause-2: It takes a really long , slow and dreary time to dope out what TUCK EVERLASTING is about .

The following are the points we noted in the guidelines:

- We asked our annotators not to think of words or emotion clauses out of context, rather they should think about them within the context for sentence annotation.
- We noted to them to not annotate the sentences and clauses according to their (e.g., cultural, religious) backgrounds.
- Our annotators were free to choose any dictionaries or resources to judge the emotion in the sentences.
- We provided one example for each emotion label (e.g. "Siri does not pick my accent and drives me crazy", where the emotion label is *anger*.).

Annotation Stage: we set up the annotation job in CrowdFlower,[1] an online crowdsourcing platform. We separate the setup for sentence level annotation from clause

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

Dataset	joy	trust	anti	surprise	sad	fear	anger	disgust	other-emo	no-emo
sentence-level										
HLN	106	6	56	31	83	68	28	55	0	662
BLG	689	43	260	150	312	132	192	255	13	2051
MOV	4875	26	119	255	258	63	20	5145	13	1081
SBHM (total)	5670	75	435	436	653	263	240	5455	26	3794
clause-level										
HLN	93	1	13	7	35	12	14	45	1	1081
BLG	1138	28	278	81	291	148	258	831	16	3665
MOV	8772	26	126	130	228	154	63	9651	9	2743
CBHM (total)	10003	55	417	218	554	314	335	10527	26	7489

Table 1: Multi-genre corpus consists of three genres and the distribution of emotion categories per sentence and clauses. Category joy and disgust are notable in movie review.

level, due to differences in task objective and slight differences in the guidelines. As such, to set up the two annotation jobs we took the following steps:

- We used the emotion categories developed in the pilot stage.
- We simplified the guidelines, which we used at the pilot stage. The only factor we noted to the annotators in the simplified guidelines was to not take emotion words or expression out of context for sentence annotation.
- We provided one example for each emotion label.
- We mixed the three datasets together and put every 5 sentences/clauses in one HIT with a compensation of $0.07 (7 cents). [2]

We provided 5000 single clause sentences annotated in sentence-level task as gold labeled data for clause-level annotation. We excluded the remaining single-clause sentences from clause-level annotation.

3.3. Annotation Evaluation

Each sentence/clause was annotated by 3 annotators. Crowdflower platform assigns a 'trust' score per annotation task. This score is a number between 0 and 1, and it is defined by the system as the accuracy score of an annotator. We required that only judgments with trust score above 0.7 are accepted. The system calculates 'trust' as follows: each HIT contains one gold item, the trust score is the percentage of correct answers to gold items. Judgments from annotators with score being below the threshold are tainted. To demonstrate the agreement among our annotators, we calculate per emotion tag, per datapoint, the number of judges who agreed on the emotion tag. We call this metric agreement class category (ACC). In our tasks, we asked for 3 judgments per datapoint and agreement of a minimum of two judges. Table 1 shows the statistics of the annotated corpora per emotion. We note that a significant number of units (sentences and clauses) are tagged with anticipation, even more than a basic EK6 emotion such as *surprise* which validates our choice of PL8 as a tagset. Table 2 shows the ACC in the annotated corpora per emotion label. The results show that on average, we achieved 79.95% IAA on sentence-level and 62.74% IAA, on clause-level, where at least two judges agreed on the emotion label per item. Table 3 presents the statistics on the EK6 tags of the original previous annotation on the HLN sub corpus as well as the BLG sub corpus of BHM and our current annotations on the sentence level. We report the HLN Pearson correlation as reported by the authors of the HLN annotated corpus and Kappa statistic on the BLG corpus. We note that the ACC values in Table 3 are different from Table 2 since these exclude statistics from the MOV corpus. For HLN, despite the fact that the two metrics are different, ACC and Pearson correlation, we note that the ACC metric is higher per emotion label in our annotation setting. We note the same trend for the BLG corpus comparing ACC metric and the Kappa statistic except for the emotion label *joy*. Table 4 shows the confusion matrix between the various labels of both HLN and BLG. We note that our IAA using crowdsourcing for only the 6 basic Ekman emotions (EK6) for BLG is 78.93% compared to the original of 76% in lab annotators in the original data set. Likewise for the HLN data set, we achieve an IAA of 93.16% with EK6 using crowdsourcing compared to 53.67% in the original annotated data set. This proves the feasibility of using crowdsourcing effectively for the task. Moreover our annotation with 10 tags (PL8) achieves an overall IAA of 76.5% for BLG and 95% for HLN. This suggests that PL8 is an appropriate level of tagging. Observe that agreement among the workers in CrowdFlower is higher than what we achieved in

Emotion	ACC≥2%	
	sentence	**clause**
joy	93.03	93.82
trust	65.33	23.64
anticipation	80.23	52.04
surprise	82.80	56.88
sadness	76.11	66.25
fear	70.34	72.29
anger	63.75	68.36
disgust	**97.32**	94.64
other-emotion	26.92	0.00
no-emotion	63.78	99.52
IAA	79.95	62.74

Table 2: The ACC≥2 percentage agreement per emotion label where at least 2 annotators agreed on the same label in the BHM corpus.

[2]We borrowed the expression HIT (Human Intelligence Task) from Amazon Mechanical Turk https://www.mturk.com. In Amazon Mechanical Turk a HIT is defined as: a question that needs an answer. A HIT represents a single, self-contained task a Worker can work on, submit an answer, and collect a reward upon completion.

the pilot stage. In pilot stage, the annotators received significant instruction and we had the opportunity to discuss different aspects of the task, while in CrowdFlower we do not have knowledge about the annotators background and we are not able to connect with them. Despite these issues, we achieve a very high general IAA on the sentence level verifying that crowdsourcing is an appropriate manner to curate annotations for emotion tags. In addition, emotion tags *trust, anticipation, fear, anger,* and *sadness* are controversial. Particularly, we received a high volume of feedback for emotion tags *fear, anger,* and *sadness*, indicating that these emotion tags are confusing, interchangeable, or can be used together for tagging data points.

Table 3 presents the comparison between the emotion la-

HLN Emotion	ACC≥2%	Pearson
joy	98.11	59.91
surprise	93.55	36.07
sadness	95.18	68.19
fear	95.59	63.81
anger	89.29	49.55
disgust	87.27	44.51
avg.	93.16	53.67
BLG Emotion	ACC≥2 %	Kappa
joy	73.00	0.77
surprise	79.33	0.60
sadness	73.72	0.68
fear	79.55	0.79
anger	69.27	0.66
disgust	94.51	0.67
avg.	78.23	0.76

Table 3: Comparing the inter-agreement we achieved with HLN & BLG datasets. In both datasets our annotation achieved higher IAA results.

Emo/dataset	joy	trust	anti	surp	sad	fear	anger	disg	other-emo	no-emo
BLG										
joy	**81.4**	1.4	5.2	2.0	1.6	0.5	0.1	1.0	0.1	6.1
surprise	21.1	0.0	0.0	**47.4**	5.9	2.5	1.6	1.6	0.0	19.4
sadness	3.3	1.1	3.3	0.5	**74.3**	3.3	1.1	8.9	0.0	4.4
fear	2.5	0.0	0.8	5.9	4.2	**65.8**	2.5	5.9	0.8	11.1
anger	1.6	0.5	1.6	3.2	9.2	3.8	**53.5**	17.4	0.0	8.7
disgust	1.1	0.0	2.8	0.5	10.9	1.7	28.3	**47.9**	0.0	5.7
no-emo	7.2	1.1	7.7	2.4	4.3	1.1	1.3	3.8	0.3	**70.3**
HLN										
joy	**21.8**	0.6	10.9	4.0	0.9	0.5	0.3	0.0	0.0	61.3
surprise	1.0	0.0	1.0	**9.7**	0.0	0.0	1.0	3.2	0.0	83.6
sadness	0.5	0.5	0.5	0.0	**23.6**	4.7	2.3	8.2	0.0	59.1
fear	0.0	0.0	1.9	1.3	7.8	**26.1**	3.2	5.2	0.0	54.2
anger	0.0	1.5	1.5	1.5	3.0	3.0	**13.6**	3.0	0.0	72.7
disgust	0.0	0.0	0.0	3.8	3.8	0.0	7.6	**34.6**	0.0	50.0
no-emo	0.5	0.0	0.0	0.5	2.3	1.1	1.7	4.6	0.0	**83.2**

Table 4: Confusion matrix for different emotion labels on the sentence level in BLG & HLN datasets of the BHM corpus and the original tags.

bels in BHM and the previous tags using EK6 data set on the sentence level for BLG and HLN. We consider the original tags (row entries) as gold. We note that the overlap between the previous annotation of BLG and our current annotation is higher than the overlap with HLN. Emotion tag *anger* is commonly confused with *disgust* compared to the number of annotations for *anger*. Observe that majority of confusion is in *no-emotion* tag. We also note that 8% of the BLG sentences, which previously were annotated as *no-emotion*, are tagged with *trust* and *anticipation*.

3.4. Emotion Tagging Difficulties in the Corpus

Manually annotating emotion data is a challenging task, due to different evaluation of emotion situations by humans. According to appraisal theory (Öhman, 1999), emotions are extracted from evaluations of events that could trigger different reactions by different people. In our annotation setting our annotators could choose one emotion tag among PL8 and *no-emotion, and other-emotion*, which can be challenging and confusing. During the annotation process, we observed that annotators are confused when they have to pick one of the {*anger, disgust, fear*} or {*trust, joy, anticipation*}. As a result, we had high number of tainted annotations during annotation stage.

Below we observe annotation tags provided for three examples from movie review corpus (MOV):

(a) "Engagingly captures the maddening and magnetic ebb and flow of friendship."

(b) "Rabbit-Proof Fence will probably make you angry."

(c) "Closings and cancellations top advice on flu outbreak."

All three sentences were annotated by 4 annotators per sentence (1 annotator vote was tainted).

Sentence (a): 2 annotators tagged that sentence as *joy*, 1 tagged it as *trust*, and 1 tagged it as *no-emotion*. While the expression "flow of friendship" triggers *trust*

Sentence (b): 2 annotators tagged it as *anger*, 1 as *anticipation*, and 1 as *disgust*. Sentence (c): 2 tagged it as *no emotion*, 1 as *fear* and another as *disgust*.

4. EDC Systems Experiment Setup and Results

For classification we devise the same experiments for tagging on both granularity levels: sentence and clause levels. We have 9 classes in our data, the PL8 and *no-emotion* [3]. We split the data to (80%,10%,10%) for training, dev, test, respectively.

Supervised model: we build our model using LIBLINEAR (SVM family) in WEKA classifiers,[4]. SVM has been applied with success to emotion classification in the literature (Aman and Szpakowicz, 2007; Mishne, 2005; Yan, 2014; Das and Bandyopadhyay, 2010; Mohammad, 2012; Özbal and Pighin, 2013). We experimented with other classifiers such as Naive Bays, Decision Tree, and Random Forest, and LIBLINEAR produced better results. We build our model combining number of features like: n-gram, POS, syntactic features like presence of adjective, adverbs, or negation (syn). To show the impact of clauses in sentence-level classification we created a feature based on clause emotion tags pattern, we refer to this feature as subordinate clauses (scla). For this feature, we study the distribution of clauses emotion tags in multi-clausal sentences. We note that the majority of those sentences with multiple clauses tend to have clauses with specific emotion labels (e.g. sentence emotion tag *joy*, have clauses with tags {*trust, anticipation, no-emotion, and surprise*}). We model this feature as an 8-dimension vector, where

[3] Authors release the dataset for research purposes upon the requests.

[4] http://www.cs.waikato.ac.nz/ ml/weka/

Emo-tag	NLH		BLG		MOV		SBHM	
	LIB	RULE	LIB	RULE	LIB	RULE	LIB	RULE
joy	92.3%	42.1%	66.2%	77.3%	85.5%	89.9%	83.4%	87.4%
trust	0	33.3%	0	28.7%	33.3%	0	13.3%	20.0%
anti	40.0%	26.3%	41.9%	35.9%	11.8%	42.8%	34.3%	37.2%
surprise	40.0%	28.5%	37.5%	61.5%	29.3%	29.4%	32.3%	36.7%
sadness	28.6%	0.06%	57.6%	38.4%	45.5%	31.2%	51.3%	34.6%
fear	20.0%	0.06%	30.0%	55.5%	28.6%	70.5%	27.3%	52.3%
anger	0	0	64.7%	71.4%	0	0	57.9%	65.2%
disgust	52.2%	40.0%	53.5%	58.8%	85.8%	92.4%	83.1%	89.7%
no-emotion	83.5%	80.9%	78.8%	80.6%	52.5%	58.0%	72.9%	75.4%

Table 5: EDC LIBLINEAR and RULEBASE f-score for each motion tags. We trained LIBLINEAR on SBHM train corpus and evaluated the system on different genre and SBHM test sets. Emotion tags with f-score of "0" are low populated categories (i.e. from 0-4 data points in the corresponding set)

each dimension represent one emotion tag with a binary value: 1 indicates the presence of sub-sentential emotion clause tag and 0 otherwise. We train LIBLINEAR model with training set and tune our parameters on dev set. Evaluation is done on test set.

Rule-base (RULEBASE): for sentence-level classification we chose one of the clause tags as sentence emotion tag. Our rule is as follow: if there is a match between one of the clause tags and sentence tag our algorithm picks that clause tag as sentence tag, if there is no match, one of the tags are selected randomly. We used SBHM training set to define this rule. This method is evaluated on SBHM test set.

Table 5 shows the results. We can observe the impact of subordinate clauses (scla) feature in supervised setting. This feature increases the system accuracy and f-score by 4.1%. Rule-based model creates f-score of 80.4% and the best results for sentence-level classification. These two results indicate the significance of clause-level annotation in sentence-level classification. Further, clause-level supervised system has a great improvement compare to baseline. **Comparisons to other systems** - we

Features	Clause		Sentence	
	acc.%	f-score%	acc.%	f-score%
Baseline (presence of emotion words)	46.2%	45.3%	47.3%	46.2%
LIBLINEAR	**71.3%**	**70.9%**	72.2%	71.3%
LIBLINEAR+scla	-	-	76.4%	75.7
RULEBASE	-	-	**81.1%**	**80.4%**

Table 6: EDC LIBLINEAR results using different combination of features on both clause and sentence levels and RULEBASE using rule-base algorithm.

compare EDC and RULEBASE on PL8 and *no-emotion* with previously reported results on two sets, i.e. NLH and BLG. However, we only compare our results with systems, which reported their results on EK6. This comparison is on sentence-level, since, clause-level emotion system is initiated in this work. Table 7 shows the comparison of LIBLINEAR and RULEBASE with other reported systems: Aman (Aman and Szpakowicz, 2007), SEMEVAL 2007 (Strapparava and Rada, 2007), Ghazi (Ghazi et al., 2010) [5], Mohammad (Mohammad, 2012), Özbal (Özbal and Pighin, 2013). We observe that RULEBASE outperforms other results for BLG, and both of our systems outperform other results for NLH. This indicates a) clause-level annotation improves sentence-level classification; b) PL8 is a better reflection for both NLH and BLG sets.

Method	Corpus	
	NLH	BLG
	acc.%	acc.%
Aman	-	73.8%
SEMEVAL 2007	17.5%	-
Ghazi	57.4%	61.6%
Mohammad	52.4%	31.4%
Özbal	20.7%	43.6%
LIBLINEAR	**74.5%**	66.1%
RULEBASE	69.7%	**75.5%**

Table 7: Comparing EDC: LIBLINEAR and RULEBASE results with previously reported results on two NLH and BLG sets. EDC and RULEBASE results are on PL8 and *no-emotion*. SEMEVAL 2007 reported results only on NLH, Aman collected BLG and reported their results only on BLG.

5. Conclusion and Future Direction

Unified annotation and combination of different genre datasets can improve and generalize emotion detection in sentences. We demonstrated that PL8 emotion tags represent these dataset better than EK6 emotion tags and if we aim to expand the emotion tagset to more fine-grain, PL8 annotation enables us to fulfill this aim. Our results showed clause-level feature can improve the prediction of emotion in sentence-level. We provide an automated system for clause-level emotion detection and classification. Further, we annotated emotions in clause-level. In future, our aim is to create sophisticated Deep Neural Network models for sentence-level classification, leveraging clause-level emotion tags. We aim to build systems that can tag smaller piece of text (i.e. phrases, clauses, words) automatically. And, we intend to add different genres to our corpus, mainly our aim is to add genres with different syntax from the current collections.

6. Acknowledgements

This study is supported by DARPA LORELEI Grant. Ali Seyfi and Sardar Hamidian carried out annotation procedure at pilot stage.

[5] They reported two different results, one is flat classification and the other is hierarchical classification. Flat classification is comparable to EDC.

7. Bibliographical References

Abdul-Mageed, M. and Lyle, U. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. *ACL*.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. *10th International Conference on Text, Speech, and Dialogue*.

Das, D. and Bandyopadhyay, S. (2010). Identifying emotional expressions, intensities and sentence level emotion tags using a supervised framework. In *PACLIC*, volume 24, pages 95–104.

Dodds, P. and Danforth, C. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11.4:441–456.

Ekman, P. (1992). An argument for basic emotions. *Cognition and emotion*, 6.3-4:169–200.

Ghazi, D., Inkpen, D., and Szpakowicz, S. (2010). Hierarchical versus flat classification of emotions in text. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 140–146. Association for Computational Linguistics.

Kim, S. M., Valitutti, A., and Calvo, R. A. (2010). Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. Association for Computational Linguistics.

Klein, D. and Manning, C. (2003). Proceedings of the 41st annual meeting on association for computational linguistics. *Association for Computational Linguistics*, Volume 1.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Mishne, G. (2005). Experiments with mood classification in blog posts. *ACM SIGIR*.

Mohammad, S. (2012). Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics.

Öhman, A. (1999). Distinguishing unconscious from conscious emotional processes: Methodological considerations and theoretical implications. *Handbook of cognition and emotion*, pages 321–352.

Özbal, G. and Pighin, D. (2013). Evaluating the impact of syntax and semantics on emotion recognition from text. In *Computational Linguistics and Intelligent Text Processing*, pages 161–173. Springer.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *43rd Annual Meeting on Association for Computational Linguistics*.

Plutchik, R. (1962). The emotions: Facts, theories, and a new model. *New York: Random House*.

Strapparava, C. and Rada, M. (2007). Semeval-2007 task 14: Affective text. *4th International Workshop on Semantic Evaluations. Association for Computational Linguistics*.

Vo, B. and Collier, N. (2013). Twitter emotion analysis in earthquake situations. *International Journal of Computational Linguistics and Applications*, 4.1:159–173.

Yan, J. (2014). Expanding the range of automatic emotion detection in microblogging text. *EACL*.