Supplemental Material

Crowdsourcing and Validating Event-focused Emotion Corpora for German and English

Enrica Troiano, Sebastian Padó and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

A Corpus Generation and Labelling

For experimental reproducibility, we detail here our crowdsourcing approach. Figure 1 illustrates the instructions presented to the annotators for sentence generation (Phase 1), Figure 2 shows a preview of the task itself. The labelling task of Phase 2 is presented in Figure 3.

To built deISEAR, we targeted Figure-Eight contributors from Germany and Austria, while the English experiment was restricted to United Kingdom and Ireland. As a quality check, we required all workers to be level-3 contributors, i.e., the most experienced ones, who reached the highest accuracy in previous Figure-Eight jobs. It should be noted that these laypeople received only minimal and distant training, while participants of ISEAR were directly instructed by the experimenters. We aimed at adapting their questionnaire to a crowdsourcing framework, by formulating the task of sentence generation as one of sentence completion (e.g. "*Ich fühlte Freude, als/weil/...*", "*I felt Joy when/because ...*"). Preliminary experiments showed that people provided more coherent and grammatically correct sentences than when they were presented with a faithful translation of the original survey.

Phase 1 involved 121 English jobs and 116 German jobs after filtering unacceptable answers (e.g. nonsensical items), totalling 2002 tasks (hits). The two languages required a diverse amount of jobs because ungrammatical and nonsensical descriptions were (manually) discarded. In the second Phase, 34 jobs were launched for English and 23 for German. This way we collected 5005 annotations for each language (i.e. 5 annotations per description). Overall, data collection and annotation was finalized in three months. The total cost was 300$ for Phase 1, and 150$ for Phase 2.

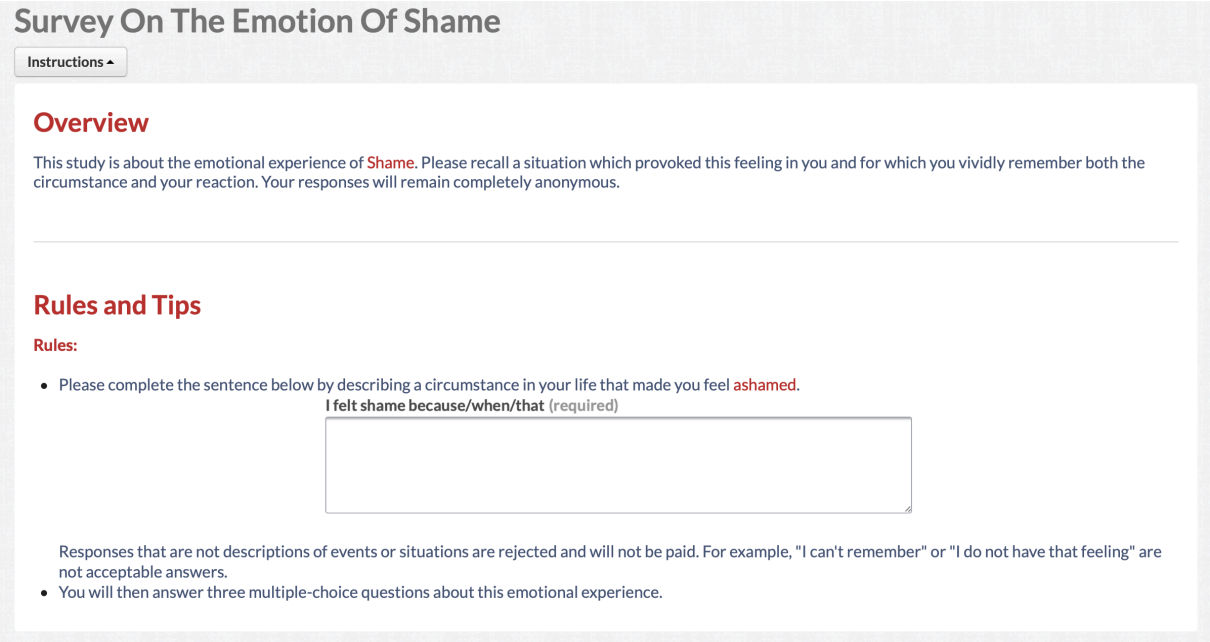

Figure 1: Instructions for the Generation Task

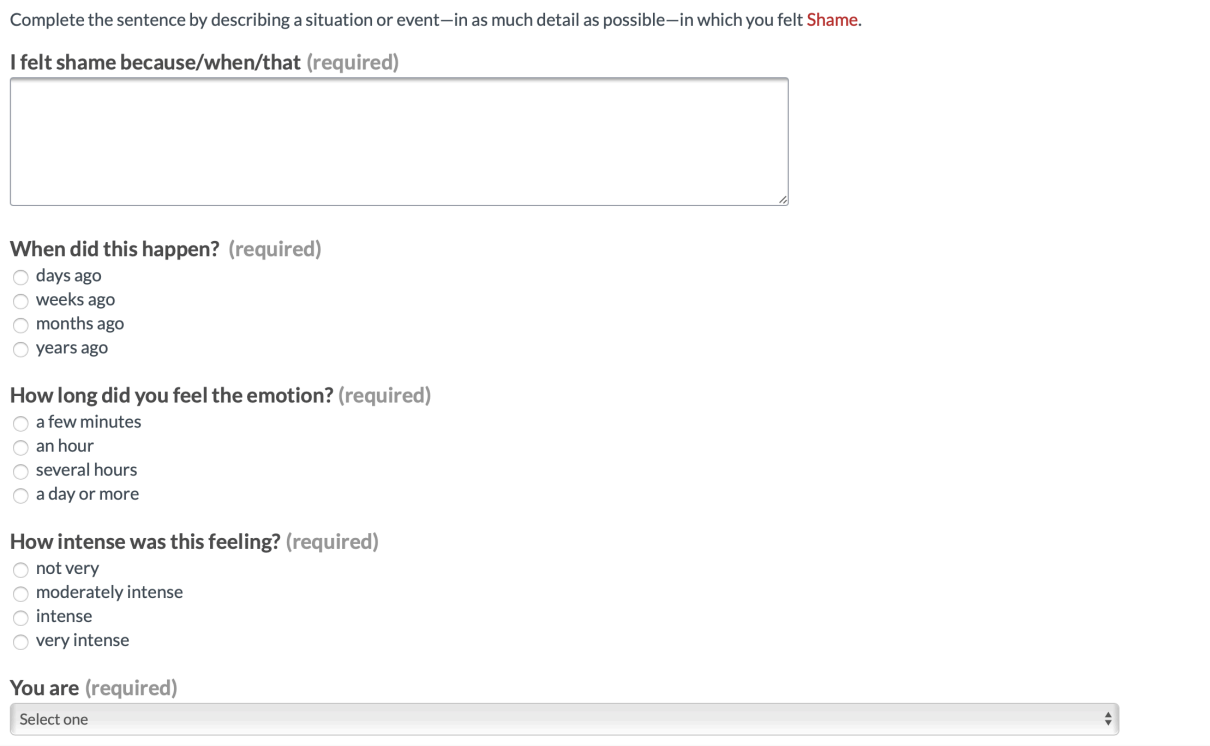

Figure 2: Preview of the Generation Task

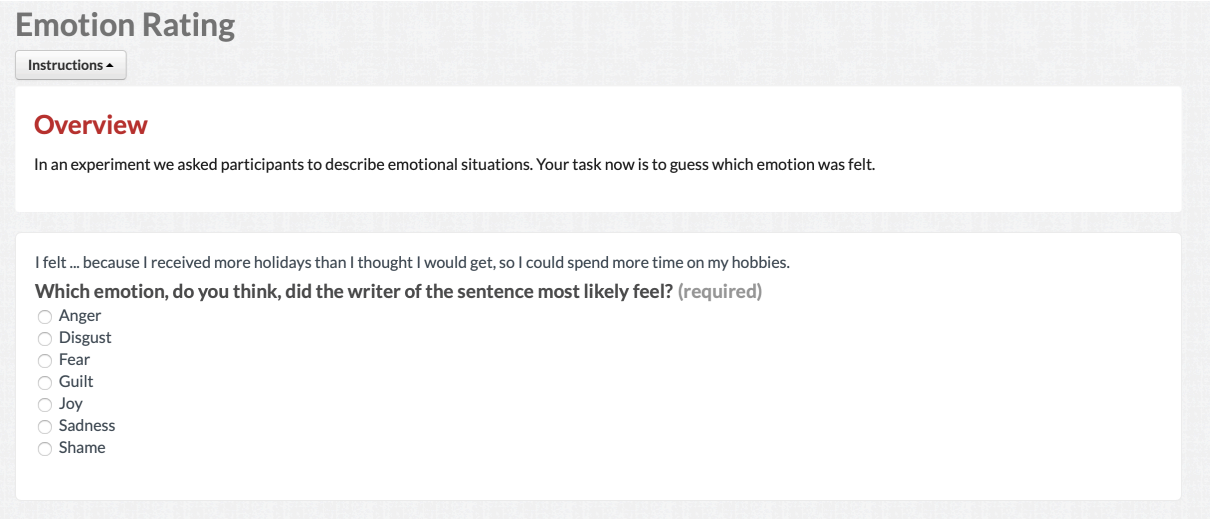

Figure 3: Preview of the Emotion Validation Task

B Descriptive Analysis

Table 1 and Table 2 present a compact description of the corpora, normalizing the counts by column and by row blocks, as reported in Section 4 in the main paper.

Table 1 highlights differences in the distribution of emotions across different temporal distances, intensities, durations, and annotators' gender. We see for instance that Shame is outstanding in English for long-distant events, while Anger and Disgust (depending on language) are more dominant in events that happened a few days prior to description production. For intensities, the distribution across emotions is most unbalanced for the label "Not Very"; for duration, Disgust is the prevailing emotion among those which lasted only a few minutes, while it is the less frequent among those which persisted for one or multiple days. The exact opposite holds for Joy and Sadness, which appear to be more durable states.

Table 2 highlights differences in the distribution of extra-linguistic labels across different emotions. A few commonalities emerge between the two languages. The majority of descriptions are referred to remote emotion episodes. Moreover, Anger-, Fear-, Joy- and Sadness-related descriptions are mostly about events which caused very intense affective states. For duration, most occurrences of Anger and Sadness lasted longer than one day both in German and English, while Fear episodes are more short-termed, similar to Disgust.

		Temporal Distance				Intensity				Duration				Gender		
	Emotion	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	Anger	.19	.12	.13	.13	.05	.10	.18	.15	.07	.16	.18	.18	.14	.14	0
	Disgust	.16	.18	.17	.08	.21	.20	.13	.10	.31	.20	.04	.01	.14	.15	0
	Fear	.10	.16	.15	.16	.07	.09	.16	.18	.16	.18	.14	.10	.14	.16	0
	Guilt	.15	.13	.12	.16	.14	.22	.15	.08	.13	.16	.20	.10	.15	.12	0
	Joy	.17	.15	.12	.14	.04	.07	.16	.20	.05	.10	.20	.23	.14	.16	1
	Sadness	.12	.13	.17	.15	.05	.12	.12	.21	.05	.05	.13	.31	.14	.14	0
	Shame	.10	.14	.15	.17	.43	.21	.11	.07	.23	.15	.11	.06	.15	.12	0
English	Anger	.18	.16	.13	.11	.12	.12	.15	.16	.13	.13	.16	.15	.15	.14	0
	Disgust	.23	.14	.11	.10	.16	.18	.12	.13	.28	.15	.11	.07	.13	.15	0
	Fear	.08	.16	.19	.15	.03	.10	.18	.17	.22	.16	.15	.07	.16	.13	0
	Guilt	.13	.14	.14	.15	.33	.18	.14	.07	.11	.22	.12	.14	.14	.15	0
	Joy	.13	.14	.16	.14	.03	.09	.15	.20	.06	.07	.19	.20	.14	.14	0
	Sadness	.16	.14	.16	.12	.13	.16	.12	.16	.07	.12	.10	.23	.15	.14	0
	Shame	.09	.12	.10	.21	.21	.18	.13	.11	.12	.14	.17	.14	.13	.15	0

Table 1: Statistics normalized by column. The unnormalized counts are shown in the paper in Table 1.

		Temporal Distance				Intensity				Duration				Gender		
	Emotion	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	Anger	.32	.17	.22	.29	.02	.17	.47	.34	.16	.20	.27	.36	.78	.22	0
	Disgust	.27	.27	.29	.17	.08	.36	.34	.22	.66	.26	.06	.02	.77	.23	0
	Fear	.17	.22	.26	.34	.03	.17	.41	.40	.35	.22	.22	.21	.76	.24	0
	Guilt	.25	.19	.21	.35	.06	.40	.38	.17	.29	.20	.30	.21	.81	.19	0
	Joy	.28	.21	.20	.31	.01	.13	.42	.44	.10	.13	.29	.48	.75	.24	.01
	Sadness	.20	.18	.29	.32	.02	.22	.30	.46	.11	.06	.19	.64	.79	.21	0
	Shame	.17	.20	.25	.38	.17	.39	.29	.15	.50	.20	.17	.13	.81	.19	0
English	Anger	.31	.20	.17	.31	.06	.24	.34	.36	.21	.16	.25	.38	.43	.57	0
	Disgust	.40	.17	.15	.28	.08	.36	.26	.30	.46	.19	.17	.18	.40	.60	0
	Fear	.13	.20	.25	.41	.01	.21	.40	.38	.36	.20	.24	.19	.46	.54	0
	Guilt	.23	.17	.19	.41	.17	.36	.30	.16	.18	.27	.20	.35	.41	.59	0
	Joy	.22	.17	.22	.39	.01	.19	.34	.46	.10	.09	.30	.51	.42	.58	0
	Sadness	.28	.17	.22	.34	.07	.31	.27	.35	.12	.15	.16	.57	.43	.57	0
	Shame	.15	.15	.13	.57	.11	.36	.29	.24	.20	.17	.27	.35	.40	.60	0

Table 2: Statistics normalized by partial row. The unnormalized counts are shown in the paper in Table 1.

C Event-type Analysis

The event-type analysis presented in Section 4 targeted 385 items per language (55 descriptions per emotion). Table 2 in the paper shows the counts of instances associated to the psychological labels across the seven emotions.

For each description, we annotated the following boolean variables:

- About the event time:
  - Does the text describe a *general event*?
  - Does the text describe a *future event*?
  - Does the text describe a *past event*?
- About the realization of the emotion:
  - Is it an actual or a *prospective* emotion?
- About the embedding in a social environment:
  - Are other people or animals part of the event description; is it a *social* event description?
- About the consequences of the event:
  - Are there *self-consequences*?
  - Are there *consequences for others*?
- About the control of the writer:
  - Is the author presumably under *situational control*?
  - Does the author presumably have *self control/responsibility*?

While the paper describes the distribution of labels by emotion, here we expand the discussion to the extra-linguistic information collected in Phase 1. Table 3 distributes the raw counts across the annotation values. It should be noticed that the random descriptions used for this analysis were not balanced with respect to their values of each variable. For this reason, Table 4 reports relative counts (i.e. counts of descriptions normalized by the number of instances within the label Day, Week, Month etc.).

Some regularities can be observed cross all columns of Table 4. For instance, events which involved a purposeful participation of their experiencer are a minority in both languages (Sit. control), and approximately 50% of the descriptions mention individuals other than the writer (Social). The latter proportion, however, is higher for English than for German.

Events that are linked to consequences for the self mostly come from the German sample (Self conseq.). In German, moreover, such type of events are recalled more frequently than events that had consequences on others (Conseq. oth.). The opposite is true for English: emotions of English authors often wrote about events that affected the life of other people or animals. This holds irrespective of the temporal distance, the intensity, the duration of the experience and the gender of the experiencer. Exceptions are English descriptions of facts which only lasted a few minutes, and which appear to bring consequences for the self more than for others (Self conseq. and Conseq. oth. in column min).

As for the responsibility of events, this label is consistent across all columns in the German sample. Instead, in English we observe some marked differences. Emotions with a low intensity (column NV) followed an event which was directly triggered by their experiencer, but very intense emotions are less frequently associated to responsibility (column VI). Lastly, shorter events (min) imply the responsibility dimension more than long ones ($\geq$d).

		Temporal Distance				Intensity				Duration				Gender		
	Dimension	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	General Event	2	3	1	2	0	1	4	3	4	0	1	3	6	2	0
	Future Event	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0
	Past Event	98	76	101	101	22	92	141	121	121	66	83	106	287	89	0
	Prospective	3	2	2	0	0	3	3	1	2	2	1	2	5	2	0
	Social	55	41	53	51	13	43	80	64	70	32	42	56	152	48	0
	Self conseq.	54	45	70	67	15	52	94	75	74	36	59	67	176	60	0
	Conseq. oth.	42	30	34	41	10	35	54	48	52	25	28	42	110	37	0
	Sit. ctrl.	17	13	18	18	2	17	29	18	21	10	14	21	56	10	0
	Responsib.	53	37	63	55	11	57	76	64	68	40	45	55	160	48	0
	Sum	226	171	242	234	51	208	341	273	291	145	191	246	666	207	0
English	General Event	6	2	2	1	2	2	5	2	5	2	1	3	3	8	0
	Future Event	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Past Event	88	61	73	152	23	104	122	125	76	74	85	139	155	219	0
	Prospective	3	4	3	5	0	2	8	5	5	4	3	3	7	8	0
	Social	73	51	56	107	14	72	94	107	49	52	66	120	103	184	0
	Self conseq.	46	30	34	90	14	57	71	58	52	32	47	69	89	111	0
	Conseq. oth.	51	38	39	73	8	49	69	75	30	47	40	84	73	128	0
	Sit. ctrl.	15	17	16	42	12	30	25	23	21	19	17	33	40	50	0
	Responsib.	50	36	47	89	20	71	80	51	57	50	53	62	104	118	0
	Sum	244	178	197	407	70	283	352	321	219	206	227	374	419	607	0

Table 3: Event-type analysis: Raw counts of the labels which were manually assigned to a subset of enISEAR and deISEAR, across the extra-linguistic information collected in Phase 1. See the text for the explanation of variables.

		Temporal Distance				Intensity				Duration				Gender		
	Dimension	D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	General Event	.02	.04	.01	.02	0	.01	.03	.02	.03	0	.01	.03	.02	.02	0
	Future Event	0	0	.01	0	0	0	.01	0	0	0	.01	0	0	0	0
	Past Event	.98	.96	.98	.98	1	.99	.97	.98	.97	1	.98	.97	.98	.98	0
	Prospective	.03	.03	.02	0	0	.03	.02	.01	.02	.03	.01	.02	.02	.02	0
	Social	.55	.52	.51	.50	.59	.46	.55	.52	.56	.48	.49	.51	.52	.53	0
	Self conseq.	.54	.57	.68	.65	.68	.56	.64	.60	.59	.55	.69	.61	.60	.66	0
	Conseq. oth.	.42	.38	.33	.40	.45	.38	.37	.39	.42	.38	.33	.39	.37	.41	0
	Sit. ctrl.	.17	.16	.17	.17	.09	.18	.20	.15	.17	.15	.16	.19	.19	.11	0
	Responsib.	.53	.47	.61	.53	.50	.61	.52	.52	.54	.61	.53	.50	.54	.53	0
English	General Event	.06	.03	.03	.01	.08	.02	.04	.02	.06	.03	.01	.02	.02	.04	0
	Future Event	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Past Event	.94	.97	.97	.99	.92	.98	.96	.98	.94	.97	.99	.98	.98	.96	0
	Prospective	.03	.06	.04	.03	0	.02	.06	.04	.06	.05	.03	.02	.04	.04	0
	Social	.78	.81	.75	.70	.56	.68	.74	.84	.60	.68	.77	.85	.65	.81	0
	Self conseq.	.49	.48	.45	.59	.56	.54	.56	.46	.64	.42	.55	.49	.56	.49	0
	Conseq. oth.	.54	.60	.52	.48	.32	.46	.54	.59	.37	.62	.47	.59	.46	.56	0
	Sit. ctrl.	.16	.27	.21	.27	.48	.28	.20	.18	.26	.25	.20	.23	.25	.22	0
	Responsib.	.53	.57	.63	.58	.80	.67	.63	.40	.70	.66	.62	.44	.66	.52	0

Table 4: Event-type analysis: Counts are normalized by instances with the particular value, e.g., the count in the cell "Time General"–"D" is normalized by the number of all instances with the associated value D (temporal distance of days).

D Annotator Agreement

Section 4.1 discussed the agreement reached by different subsets of annotators *at each generation label*. We report relative counts in Table 5 and we extend the analysis in Table 6, summing over the prompting emotions. This table shows the interannotator agreement of Phase-2 annotators with respect to the meta-information given by the participants of Phase 1, i.e., all the alternatives for gender, intensity, duration and temporal distance under the column Labels.

These numbers represent the count of descriptions within a corpus – and *not within a generation label*, for which the annotation label is the same as the generation label. One can read the table as follows: 177 descriptions from deISEAR, which were labeled as VI by Phase 1 annotators, were then labelled by 5 Phase 2 annotators with their original prompting emotion; 506 instances provided by female annotators for enISEAR were labelled by at least 2 Phase 2 annotators with their original prompting emotion, and so on.

Notably, in the table of Section 4.2, the maximum value that each cell can reach is 143, i.e., the total number of descriptions prompted by a specific emotion. Here, the maximum value varies by cell, because each meta-data label is assigned to a different number of descriptions[1]. Accordingly, higher counts do not necessarily indicate stronger agreement.

	German					English				
Emotion	≥1	≥2	≥3	≥4	=5	≥1	≥2	≥3	≥4	=5
Anger	.94	.87	.75	.57	.36	.96	.90	.78	.62	.41
Disgust	.97	.94	.91	.87	.64	.83	.71	.59	.53	.37
Fear	.94	.87	.76	.69	.55	.95	.92	.87	.81	.60
Guilt	.96	.88	.71	.47	.22	.96	.91	.87	.62	.31
Joy	.99	.99	.99	.98	.95	1	1	1	1	.96
Sadness	.92	.86	.79	.68	.53	.98	.93	.92	.81	.68
Shame	.90	.76	.60	.46	.29	.81	.64	.45	.29	.16
Sum	6.62	6.17	5.51	4.71	3.53	6.48	6.01	5.47	4.69	3.49

Table 5: Relative agreement counts.

		German					English				
	Labels	≥1	≥2	≥3	≥4	=5	≥1	≥2	≥3	≥4	=5
When	D	226	157	184	209	226	229	211	189	161	115
	W	197	184	169	143	108	168	152	137	112	79
	M	229	215	198	174	125	177	165	154	138	109
	Y	295	275	237	200	161	353	331	302	259	196
Length	min	291	275	245	213	145	223	208	185	162	115
	h	173	162	151	127	99	162	145	130	106	74
	>h	205	188	164	139	103	210	197	178	158	118
	≥d	278	258	228	195	158	332	309	289	244	192
Intense	NV	52	46	38	32	18	74	69	61	51	31
	M	241	224	194	162	113	264	240	217	185	128
	I	352	331	301	255	197	288	267	247	213	165
	VI	302	282	255	225	177	301	283	257	221	172
Gender	M	738	684	604	510	392	386	353	316	273	200
	F	208	198	183	163	112	541	506	466	397	299
	O	1	1	1	1	1	–	–	–	–	–

Table 6: Full agreement information for both German and English crowd-sourced corpora.

[1] For an overview of the distribution of meta-data labels over the descriptions, refer to Section 4.1.

E Modeling

Table 7 shows the results of the maximum entropy classifier across all emotions.

	deISEAR						enISEAR					
Emotion	TP	FP	FN	P	R	F1	TP	FP	FN	P	R	F1
Anger	29	30	114	.49	.20	.29	27	32	116	.46	.19	.27
Disgust	65	57	78	.53	.45	.49	67	85	76	.44	.47	.45
Fear	70	77	73	.48	.49	.48	85	69	58	.55	.59	.57
Guilt	75	140	68	.35	.52	.42	79	161	64	.33	.55	.41
Joy	106	61	37	.63	.74	.68	94	43	49	.69	.66	.67
Sadness	63	31	80	. 67	.44	.53	70	29	73	.71	.49	.58
Shame	66	131	77	.34	.46	.39	49	111	94	.31	.34	.32
Micro	474	527	527	.47	.47	.47	471	530	530	.47	.47	.47

Table 7: Classification results for both corpora.