SemEval-2023 Task 9: Multilingual Tweet Intimacy Analysis

Jiaxin Pei ♣ **Vítor Silva**† **Maarten Bos** † **Yozen Liu**†
Leonardo Neves†‡ **David Jurgens**♣ **Francesco Barbieri**†
♣School of Information, University of Michigan, Ann Arbor, MI, USA
†Snap Inc., Santa Monica, CA, USA
‡ Grammarly, San Francisco, CA, USA
♣{pedropei, jurgens}@umich.edu
†{fbarbieri, vsilvasousa, maarten, yliu2}@snap.com
‡leo.neves@grammarly.com

Abstract

Intimacy is an important social aspect of language. Computational modeling of intimacy in language could help many downstream applications like dialogue systems and offensiveness detection. Despite its importance, resources and approaches on modeling textual intimacy remain rare. To address this gap, we introduce MINT, a new **M**ultilingual **int**imacy analysis dataset covering 13,372 tweets in 10 languages including English, French, Spanish, Italian, Portuguese, Korean, Dutch, Chinese, Hindi, and Arabic along with SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. Our task attracted 45 participants from around the world. While the participants are able to achieve overall good performance on languages in the training set, zero-shot prediction of intimacy in unseen languages remains challenging. Here we provide an overview of the task, summaries of the common approaches, and potential future directions on modeling intimacy across languages. All the relevant resources are available at https://sites.google.com/umich.edu/semeval-2023-tweet-intimacy.

1 Introduction

Intimacy has long been viewed as a primary dimension of human relationships and interpersonal interactions (Maslow, 1981; Sullivan, 2013; Prager, 1995). Existing studies suggest that intimacy is an essential social component of language and can be modeled with computational methods (Pei and Jurgens, 2020). Recognizing intimacy can also serve as an important benchmark to test the ability of computational models to understand social information (Hovy and Yang, 2021).

Despite the importance of intimacy in language, resources on textual intimacy analysis remain rare. Pei and Jurgens (2020) annotated the first textual intimacy dataset containing 2,397 English questions, collected mostly from social media posts and fictional dialogues. However, such question phrases are often used primarily for interrogative situations, and, as such, models trained over the dataset may not generalize well to other types of text.

To further promote computational modeling of textual intimacy, we introduce a new **M**ultilingual textual **int**imacy dataset (MINT). The training data in MINT covers tweets in 6 languages, including English, Spanish, French, Portuguese, Italian, and Chinese, which are languages used by over 3 billion people on Earth, in The Americas, Europe, and Asia. A total of 12,000 tweets are annotated for the six languages. To test the model generalizability under zero-shot settings, we also include small test sets for Dutch, Korean, Hindi, and Arabic (500 tweets for each), which are spoken by over 0.8 billion people around the world.

We benchmarked a series of large multilingual pre-trained language models including XLM-T (Barbieri et al., 2021), XLM-R (Conneau et al., 2019), BERT (Devlin et al., 2018), DistillBERT (Sanh et al., 2019) and MiniLM (Wang et al., 2020). We found that distilled models generally perform worse than other normal models, while an XLM-R model trained over the Twitter dataset (XLM-T) performs the best on seven of the ten languages. While the pre-trained language models are able to achieve promising performance, zero-shot prediction of unseen languages remains challenging, especially for Korean and Hindi.

Based on MINT, we organized SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis, which attracted 45 teams from more than 20 countries. Participants primarily built their systems based on multilingual pre-trained language models like XLM-T (Barbieri et al., 2021), mBERT (Devlin et al., 2018), and XLM-R (Conneau et al., 2019) due to the multilingual nature of this task. However, participants also used a variety of model-building techniques to potentially improve performance, with data augmentation methods like translation, external label generation, and word-level

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2235–2246

substitution being among the most heavily used. Overall 24 out of 45 teams are able to achieve >0.7 Pearson's r on the training languages and 22 teams are able to achieve >0.4 Pearson's r on the unseen languages, which beat the strong XLM-T baseline.

2 Task description

Intimacy is one of the most fundamental dimensions of human relationships (Prager, 1995). Intimacy has long been used as the index for not only relationships but also the interactions between people (Hinde, 1981). Language plays a central role in interpersonal interactions as it allows people to communicate various types of information, share emotions and build connections (Hartley, 2002). One of the most prominent forms of interaction involving language is self-disclosure, which refers to the process of sharing personal experiences or emotions about themselves (Cozby, 1973). Due to the importance of self-disclosure in interpersonal relationships, existing studies on intimacy in language generally focus on self-disclosure. Those studies usually consider intimacy as the primary dimension of self-disclosure, where deep disclosures of personal emotions or relationships are considered as intimate (Jourard and Lasakow, 1958; Snell et al., 1988). Datasets and NLP models are also built to automatically analyze self-disclosure in communications, especially on social media (Bak et al., 2012, 2014). However, self-disclosure is only part of the language that people use in daily communications. Solely studying intimacy in the context of self-disclosure overlooks many types of language interactions which do not involve self-disclosure. For example, Pei and Jurgens (2020) build the first dataset of English questions and found that questions can also have various levels of intimacy.

While the question intimacy dataset provides resources to study intimacy in language, it only focuses on questions, missing the variation of intimacy in other types of texts. Moreover, the expressions of intimacy depend on the specific types of language. Different languages may have different expressions of intimacy. Therefore, to support further studies on modeling intimacy in different languages, we present SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis.

Multilingual Tweet Intimacy Analysis is a task to predict the intimacy of tweets in different languages. Defining intimacy in language is challenging as intimacy is a natural concept (Prager, 1995). In our task, we focus on perceived intimacy by asking annotators to give their subjective judgment of tweet intimacy. We draw a diverse pool of native speakers for each language from Prolific.co to participate in our annotation task. Each annotator is asked to answer "How intimate do you think the given tweet is?" using a 1-5 likert scale, which finally leads to 13,372 tweets in 10 languages annotated with an intimacy label.

The training data contains labeled intimacy for six languages: English, French, Spanish, Italian, Portuguese, and Chinese. Pre-trained language models (PLMs) have achieved huge advances in recent years and have shown promising zero-shot abilities for downstream tasks in different languages (Conneau and Lample, 2019; Devlin et al., 2018). To encourage new studies on understanding intimacy in language as well as understanding zero-shot ability of PLMs, we also include four other languages without training data (Dutch, Hindi, Korean, and Arabic). The participants are asked to build models using the six training languages that can predict tweet intimacy from 1 (not intimate at all) to 5 (very intimate). The final model performance is evaluated on the test set in all ten languages and Pearson's r is used as the final evaluation metric.[1]

3 Data

We choose Twitter as the source of our dataset because Twitter is a public social media platform that naturally includes multilingual text data and, from our analysis, a fair amount of intimate texts. In this section, we introduce the data collection and annotation process for MINT.

3.1 Sampling

We use tweets sampled from 2018 to 2022. We use the *lang_id* key in the tweet object to select English and Chinese tweets. For other languages, we use fastText (Joulin et al., 2016b,a) for language identification[2] and assign language labels when the model confidence is larger than 0.8. All the mentions of unverified users are replaced with a special token "@user" during pre-processing to remove noise from random and very infrequent usernames.

[1] An alternative evaluation was considered, namely using the mean r across each language's performance. However, this approach could allow models to use varying scales across languages, leading to lower comparability of the scores for texts in different languages.

[2] https://fasttext.cc/docs/en/language-identification.html

We fine-tune XLM-T, a multilingual RoBERTa model adapted to the Twitter domain (Barbieri et al., 2021) over the annotated question intimacy dataset (Pei and Jurgens, 2020). The fine-tuned model attained a Pearson's r of 0.80 for English questions[3] and we use it to estimate the intimacy of all the collected tweets in 10 languages. Then, in the second step, we split the tweets into 5 buckets based on the estimated intimacy and up-sampled relatively more intimate tweets. We did bucket sampling for 1,000 English tweets and randomly sampled another 1,000 English tweets as well as all the tweets for the rest of the languages[4].

3.2 Annotation

We recruited annotators from Prolific.co and paid them $15 USD per hour for their annotations. We set a "first language" requirement during annotator pre-screening. For example, an annotator must meet the requirement of "Spanish as the first language" to annotate the intimacy of Spanish tweets. Intimacy is annotated using a 5-point Likert scale where 1 indicates "Not intimate at all" and 5 indicates "Very intimate". The annotators are asked "How intimate do you think the given tweet is?" (translated into the corresponding language when tweets in a certain language are presented) and are encouraged to apply their own subjective judgment.

In pilot annotations, we explored 7-point likert scales as well as Best-Worst-Scaling (BWS) similar to Pei and Jurgens (2020) and calculated the Krippendorff's α to measure inter-annotator agreement (IAA). We found that the 5-point Likert scale annotations (α = 0.38) achieve IAA similar to BWS (α = 0.36) and have higher IAA than a 7-point likert scale (α = 0.25). For each language in the training set, we collected annotations for 2,000 tweets. Because they annotated the intimacy of tweets, the annotators could see sexual or potentially offensive content during annotation. Therefore, we required annotators to be at least 18 years old to work on our task. Each annotator was explicitly notified about the potential for sexual or offensive content and they signed a consent form before starting the annotations.

Each tweet was annotated by 7 annotators and each annotator was shown 50 tweets. After the annotation, each annotator was required to complete a post-study survey about their demographics including gender, age, religion, country, educational background, and occupation. For tweets that were not in the target language or that did not make sense (e.g. random characters), the annotators were instructed to annotate them as *Invalid Tweet*. We used POTATO (Pei et al., 2022) to set up all the annotation interfaces.

3.3 Quality control

Annotating textual intimacy is challenging because of the subjective nature of intimacy perception and potential individual rating bias. We designed a series of quality control procedures throughout the annotation process: (1) we conducted 10 pilot studies on Prolific.co and revised our annotation procedures according to attained IAA and participant feedback; (2) annotation guidelines for each language were carefully translated by native translators,[5] which prompts the annotators to think about intimacy in their own languages; (3) all instructions in the recruitment phase were written in the annotator's indicated first language in the recruitment phase, which could potentially remove potential spam annotators in crowdsourcing platforms; (4) we randomly inserted two attention test questions[6] to identify potential spammers; (5) the annotators were balanced by sex (based on Prolific's built-in feature) and were also generally diverse regarding other demographics (e.g. the annotators are from 73 unique countries and regions), which allowed us to collect more population-representative ratings.

3.4 Post processing

We first removed annotations from users who failed the attention test. No more than 2 annotators per language were removed in this step, except for Hindi (26 removed), Korean (7 removed), and Arabic (4 removed). To remove potential noise in the

[3] Pei and Jurgens (2020) report a Pearson's r of 0.82 using RoBERTa-base. Given that XLM-T is pre-trained on tweets, a Pearson's r of 0.8 is reasonable.

[4] We intended to do bucket sampling for all the data, however, due to an issue in the pre-processing, we were only able to do it for 1,000 tweets. We conducted further analyses for the potential effect of this error. We found that the distribution of the final annotated intimacy scores are not changed much, while the fine-tuned XLM-T only achieved a Pearson's r of 0.43 on the random sample, suggesting that the model trained on Reddit questions may not be reliable enough to detect intimacy in tweets.

[5] Chinese, Spanish, Dutch, French, Korean, Portuguese, Hindi, and Italian guidelines were translated by native speakers at Snap Inc. and the University of Michigan. The Arabic guideline was translated by one expert translator and one expert proofreader recruited from Upwork.com;both were paid $13/h.

[6] "This is a test question, please select N" where N was a random number between 1-5.

English	Intimacy
Ukrainian Railways Chief Says 'Honest' Belarusians Are Cutting Russian Supplies By Train http	1.00
19:04h Temp: 28.9°F Dew Point: 19.40°F Wind:SSW 4.3mph Rain:0.00in. Baro:29.66 inHg via MeteoBridge 3.2	1.00
A team that shops together stays together...helping life go right @StateFarm http	1.00
Leicester City fans - keep an eye on Ross Barkley. Could be moving to the Foxes on a permanent for £11m... #lcfc	1.25
@user They aren't open	1.25
Kenya I vote for #Butter for #BestMusicVideo at the 2022 #iHeartAwards @BTS_twt	1.25
That might have been the best episode of power ever	1.40
@user Coming to USA if Trump loses in 2020.	1.40
Change the formula to get a different result	1.60
@user thank u	2.50
@user Happy birthday!	2.60
it's the worst feeling when you feel like no matter how much u do for a person you'll never get the same in return	3.00
@user you're not my mom	3.00
@user @user Love you	4.00
I am SO ecstatic I'm not married to a man who has cheated on me.	4.33
My nails so mf ghetto. I'm embarrassed	4.67
need a kiss	4.75

Table 1: A sample of annotated tweets in English

crowdsourcing setting, similar to trimmed mean (Millsap and Maydeu-Olivares, 2009), we removed one highest score and one lowest score for tweets with at least five labels. After all the processing, we kept the tweets with at least two valid scores. For external test languages (i.e. Dutch, Hindi, Korean, and Arabic), we only kept tweets with a relatively low label diversity (i.e. standard deviation lower than 1) to ensure a good golden test set for the zero-shot setting[7]. The final intimacy score is calculated as the mean score of all the remaining labels for each tweet.

3.5 Annotation result

The final dataset includes 13,372 tweets annotated with the textual intimacy score. Table 2 shows the final statistics for the annotated data. We attained moderate inter-annotator-agreement, similar to previous work (Pei and Jurgens, 2020). Given the subjective nature of intimacy perception, we believe that such an IAA score is promising. To further verify the quality of the annotations, we conduct a split-half-reliability test (SHR; Johnson and Penny, 2022): randomly splitting labels into two groups and calculating the Pearson correlation between the aggregated scores from the two groups. All the SHR scores are above 0.63 with an average of 0.68, suggesting that the final aggregated scores are reliable. Figure 5 shows the intimacy distribution of the final dataset. The final dataset for English, Spanish, French, Italian, Portuguese, and

[7]40%, 15%, 17%, 16% of tweets are removed for Hindi, Dutch, Korean, and Arabic respectively.

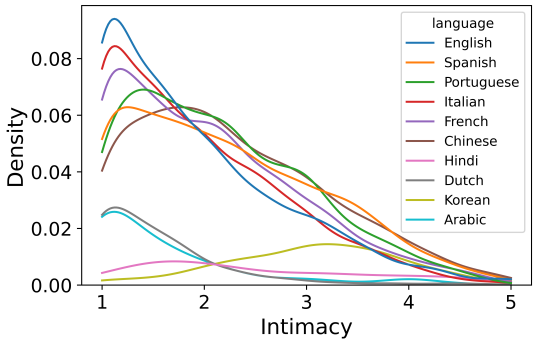

Figure 1: The distribution of intimacy scores for each language

Chinese is split into training, validation, and test sets following a ratio of 7:1:2, and all the tweets are held as the test set for Arabic, Dutch, Korean, and Hindi.

4 Baseline Models

We benchmark several baseline models on the tweet intimacy prediction task. We compare the following multilingual pre-trained language models:

1. BERT (Devlin et al., 2018): multilingual BERT model.
2. XLM-R (Conneau et al., 2019): multilingual RoBERTa model.
3. XLM-T (Barbieri et al., 2021): Multilingual RoBERTa model trained over 200M tweets.
4. DistillBERT (Sanh et al., 2019): Multilingual distilled BERT model.

Language	α	SHR	Amount
English	0.48	0.69	1,983
Spanish	0.52	0.72	1,991
Portuguese	0.45	0.66	1,994
Italian	0.43	0.63	1,916
French	0.47	0.67	1,981
Chinese	0.44	0.64	1,996
Hindi	0.61	0.68	280
Korean	0.53	0.67	411
Dutch	0.48	0.68	413
Arabic	0.58	0.74	407

Table 2: Statistics for the annotated dataset

model	XLM-T	BERT	XLM-R	DistillBERT	MiniLM
English	**0.70**	0.59	0.65	0.55	0.61
Spanish	**0.73**	0.62	0.64	0.61	0.67
Portuguese	**0.65**	0.54	0.61	0.52	0.53
Italian	**0.70**	0.57	0.67	0.58	0.62
French	**0.68**	0.55	0.63	0.54	0.57
Chinese	0.70	0.65	**0.72**	0.67	0.65
Hindi	**0.24**	0.09	0.24	0.17	0.18
Dutch	0.59	0.47	**0.60**	0.44	0.57
Korean	0.35	0.32	0.33	0.26	**0.41**
Arabic	**0.64**	0.35	0.48	0.32	0.38
overall	**0.58**	0.48	0.53	0.52	0.53

Table 3: Performance of the baselines. Hindi, Dutch, Korean, and Arabic are tested under the zero-shot setting. XLM-T achieves the best performance on 7 languages.

5. MiniLM (Wang et al., 2020): Multilingual MiniLM model.

All the models are trained with 10 epochs and the best-performing model is selected based on the dev set[8]. We train each model with 5 different random seeds and report the mean score. The learning rate is set as 0.001 and the batch size is 64. We use AdamW as the optimizer (Loshchilov and Hutter, 2017).

Table 3 shows the performance of the baselines. We found that XLM-T achieved the best performance over 7 languages, suggesting that domain-specific language model training is beneficial for our tweet intimacy analysis task.

For zero-shot tasks, while XLM-T still performs the best on Hindi and Arabic, XLM-R and MiniLM achieved the best result on Dutch and Korean, respectively. Moreover, the zero-shot performance is generally lower compared with the tasks with in-domain training, suggesting that the zero-shot task is challenging. We encourage the Task participants to explore different strategies to improve the zero-shot intimacy prediction performance.

[8]We evaluate the model performance every 500 steps and choose the best model.

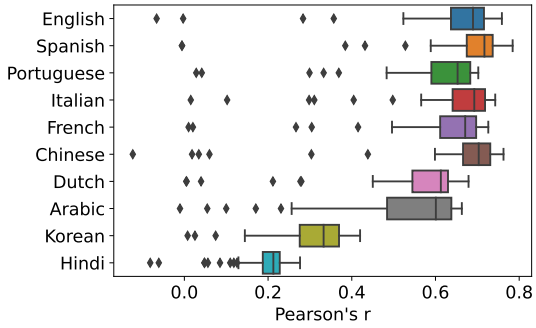

Figure 2: Overall performance on each language. The box indicates the lower quartile to the upper quartile and the whisker indicates the maximum and the minimum. Outliers are shown as dots. Participants generally achieve better performances on languages in the training set and achieved good performance on Arabic and Dutch. Predicting intimacy in Hindi and Korean remains challenging. Moreover, performances on unseen languages generally have larger variances.

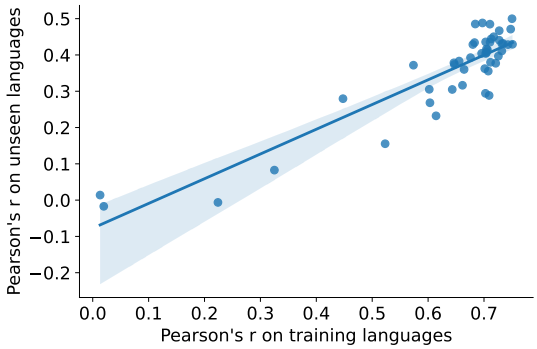

Figure 3: Models that perform better on the training languages also generally perform better on the unseen languages.

5 Results

This task received final submissions from 45 teams. Each team was allowed to submit multiple times but only the final submission was scored on the test set. Of the 45 teams, 22 submitted a system description paper. Here, we summarize the submissions' approaches by the following aspects: base pre-trained language model, data augmentation, and other heuristic techniques.

5.1 Overall performance

Among the 45 teams, 37 submissions are able to beat the M-BERT baseline and 18 participants are

Rank	Team	PLM				LMFT	Data augmentation				Ensemble	Others
		XLM-T	XLM-R	TwHin-BERT	Others		Translation	External labels	Rebalancing	Others		
1	lazybob (Yuan and Chen, 2023)	✓	✓	✓		✓	✓		✓		✓	Multi-sample Dropout, Adversarial weight perturbation, Group-layer wise learning rate decay
2	UZH_CLyp (Michail et al., 2023)	✓	✓					✓		use chatgpt to generate data samples	✓	Head-First Fine-tuning
3	opi (Dadas, 2023)		✓			✓		✓		pseudo-labeling	✓	
4	tmn (Glazkova, 2023)	✓					✓				✓	
7	DUTH (Arampatzis et al., 2023)	✓	✓								✓	
8	Zhegu (He and Zhang, 2023)		✓								✓	Exponential Penalty MSE, frozen Tuning, contrastive learning
9	arizonans (Bozdag et al., 2023)	✓										
10	Irel (Manoj et al., 2023)						✓					emoji representation with emoji2vec
11	ODA_SRIB (Kumar et al., 2023)	✓	✓				✓					Adversarial Weight Perturbation + BCE Loss
13	YNU-HPCC (Chen et al., 2023a)	✓										Bidirectional GRU on top of XLM-t
14	ZBL2W (Zhang et al., 2023)	✓					✓			Word-level: Substitution		Adversarial training + Ordinal regression
19	MaChAmp (van der Goot, 2023)				mluke-large							
27	WKU (Zheng, 2023)			✓			✓				✓	
28	HULAT (Segura-Bedmar, 2023)	✓								synonym replacement provided by EDA		
29	NLP-LISAC (Benlahbib and Boumhidi, 2023)	✓					✓	✓				
31	UMUTeam (García-Díaz et al., 2023)		✓		mBERT + mDeBERTa		✓	✓			✓	linguistic features
32	Sea_and_Wine (Chen et al., 2023b)		✓				✓		✓			Focal MSE loss
33	WADER (Suri et al., 2023)				XLNET		✓		✓	Distribution based Sampling + Difference Based Sampling	✓	Label Validation with a baseline model
34	ROZAM (Rostamkhani et al., 2023)	✓					✓					
35	ChaPat (Chavan and Patwardhan, 2023)	✓	✓		mBERT		✓	✓			✓	
37	I2C-Huelva (Pichardo Estevez et al., 2023)			✓			✓		✓			
41	jelenasteam (Lazi and Vujnovi, 2023)	✓										
45	CKingCoder (Balasubramanian et al., 2023)	✓	✓		mBERT							

Table 4: Summaries of submitted solutions. Participants generally focus on fine-tuning multilingual pre-training language models. Data augmentation and ensemble methods are also widely adopted. PLM referees to pre-trained language models. LMFT refers to in-domain language model fine-tuning (also known as continuous pre-training).

beating the strong XLM-t baseline. The best solution (LAZYBOB (Yuan and Chen, 2023)) achieves 0.616 Pearson's r on the test set. The highest single-language performance is achieved on Spanish by KING001 (Pearson's $r = 0.784$) and the overall correlation on the training languages is generally above 0.7, suggesting that the models are able to accurately predict intimacy in tweets, once fine-tuned on the labeled dataset. However, the overall performance on the unseen languages (Hindi, Dutch, Korean, Arabic) is relatively lower, especially on Hindi and Korean, which might be caused by the different distributions of the test data. Despite this, participants are able to achieve better performances on these unseen languages than the already very strong XLM-T baseline. Detailed descriptions of each system is available in the system papers. Table 5 shows the full leaderboard on this task.

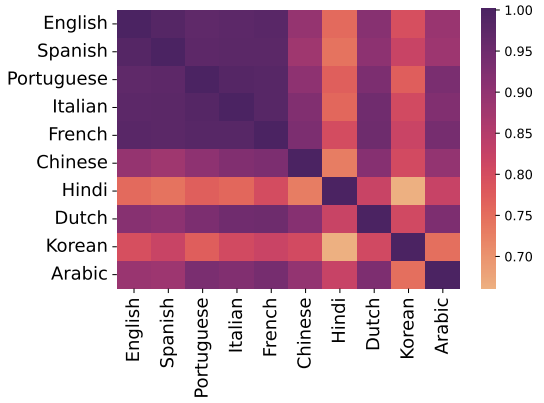

Figure 4: Correlation of model performances across languages. Teams who perform well on English, Spanish, Portuguese, Italian, and French generally perform well on all these languages. However, performance on Chinese has a lower correlation compared with other languages in the training set, potentially because Chinese is very different from the other languages.

5.2 Base pre-trained language models

Given that predicting the intimacy for unseen languages is an important part of this task and the recent advances of multilingual pretrained language models, most of the participants are using PLMs as the base architecture of their solutions. Participants primarily used three types of PLMs including XLM-t (Barbieri et al., 2021), XLM-R Conneau and Lample, 2019, and TwHin-BERT as their base PLMs. Some participants also explored m-BERT (Devlin et al., 2018), mDeBERTa, XLNET, and mluke. Some participants (e.g., JELENASTEAM (Lazi and Vujnovi, 2023)) reported experimenting with linguistic features as well as word-level representations but did not use them as the final submissions because of their low performance and difficulty to do cross-lingual tasks.

team	ranking	overall	English	Spanish	Portuguese	Italian	French	Chinese	Hindi	Dutch	Korean	Arabic
lazybob	1	**0.616**	0.758	0.770	0.689	0.739	**0.726**	0.756	0.226	0.623	**0.414**	0.643
UZH_CLyp	2	0.614	0.722	0.740	0.689	0.723	0.710	0.718	0.224	0.619	0.380	0.636
opi	3	0.613	0.749	**0.775**	**0.702**	**0.743**	0.695	**0.763**	0.238	**0.679**	0.370	**0.663**
tmn	4	0.599	0.717	0.740	0.684	0.734	0.708	0.721	0.242	0.639	0.361	0.662
OPD	5	0.599	0.728	0.746	0.699	0.735	0.701	0.734	0.223	0.640	0.333	0.652
lottery	6	0.598	0.722	0.750	0.697	0.733	0.695	0.731	0.212	0.643	0.321	0.647
DUTH	7	0.598	0.705	0.699	0.656	0.691	0.683	0.685	0.228	0.626	0.333	0.602
Zhegu	8	0.596	0.709	0.729	0.655	0.718	0.692	0.748	0.228	0.672	0.372	0.637
arizonans	9	0.594	0.674	0.735	0.661	0.727	0.707	0.711	0.259	0.601	0.339	0.658
irel	10	0.592	0.706	0.725	0.648	0.727	0.628	0.698	0.203	0.591	0.307	0.644
ODA_SRIB	11	0.589	0.716	0.747	0.702	0.733	0.708	0.735	0.213	0.642	0.369	0.636
king001	12	0.587	0.759	**0.784**	0.688	0.738	0.726	0.749	0.265	0.621	0.395	0.640
ynu_hpcc	13	0.584	0.711	0.739	0.661	0.710	0.694	0.687	0.185	0.645	0.352	0.657
ZBL2W	14	0.581	0.699	0.724	0.663	0.693	0.667	0.700	0.199	0.635	0.312	0.589
water	15	0.580	0.713	0.753	0.682	0.719	0.716	0.722	0.208	0.633	0.412	0.636
cyclejs	16	0.580	0.759	0.746	0.688	0.738	0.726	0.745	0.265	0.621	0.394	0.641
75alcool	17	0.576	0.683	0.732	0.676	0.691	0.710	0.710	0.222	0.607	0.420	0.644
MaChAmp	18	0.575	0.682	0.717	0.652	0.667	0.699	0.734	**0.276**	0.613	0.374	0.557
XLM-T		0.575	0.696	0.726	0.653	0.696	0.683	0.703	0.242	0.590	0.352	0.637
GUTS	19	0.573	0.699	0.720	0.668	0.696	0.671	0.703	0.190	0.620	0.328	0.552
CEIANLP	20	0.572	0.719	0.732	0.693	0.718	0.691	0.712	0.223	0.617	0.274	0.618
Uniretro	21	0.570	0.659	0.724	0.598	0.682	0.653	0.710	0.230	0.627	0.360	0.638
antins	22	0.569	0.669	0.677	0.600	0.656	0.671	0.733	0.207	0.616	0.368	0.601
SOJE	23	0.567	0.690	0.716	0.654	0.681	0.681	0.730	0.239	0.656	0.308	0.622
GUTS	24	0.557	0.711	0.724	0.689	0.714	0.676	0.690	0.198	0.609	0.277	0.617
UM6P_CS	25	0.557	0.704	0.697	0.624	0.669	0.620	0.677	0.223	0.627	0.395	0.581
kean_nlp	26	0.555	0.732	0.738	0.628	0.701	0.666	0.737	0.234	0.551	0.359	0.491
HULAT	27	0.551	0.722	0.698	0.699	0.693	0.670	0.710	0.210	0.642	0.257	0.601
NLP-LISAC	28	0.549	0.677	0.723	0.665	0.711	0.671	0.700	0.192	0.597	0.329	0.625
uchiha	29	0.533	0.632	0.681	0.619	0.643	0.614	0.701	0.177	0.557	0.294	0.594
UMUTeam-SINAI	30	0.532	0.642	0.705	0.582	0.659	0.611	0.704	0.220	0.539	0.362	0.503
Sea_and_Wine	31	0.529	0.658	0.638	0.599	0.666	0.613	0.649	0.217	0.550	0.211	0.495
WADER	32	0.527	0.642	0.683	0.582	0.639	0.611	0.671	0.110	0.570	0.363	0.477
ROZAM	33	0.526	0.716	0.702	0.633	0.696	0.711	0.708	0.213	0.569	0.287	0.552
chapat	34	0.519	0.683	0.716	0.664	0.702	0.689	0.736	0.205	0.624	0.224	0.544
heihei	35	0.510	0.669	0.697	0.629	0.644	0.652	0.661	0.232	0.632	0.198	0.583
I2C_Huelva	36	0.497	0.623	0.673	0.620	0.631	0.579	0.659	0.206	0.450	0.253	0.405
YNU-HPCC	37	0.488	0.541	0.589	0.483	0.498	0.540	0.678	0.213	0.452	0.386	0.424
mBERT		0.477	0.593	0.616	0.543	0.566	0.547	0.652	0.085	0.466	0.316	0.351
INGEOTEC	38	0.462	0.629	0.653	0.570	0.593	0.543	0.599	-0.061	0.280	0.230	0.401
jelenasteam	39	0.460	0.541	0.673	0.572	0.621	0.574	0.640	0.126	0.501	0.403	0.427
HappyNLP_77	40	0.367	0.605	0.617	0.507	0.611	0.496	0.438	0.129	0.485	0.332	0.256
PanwarJayant	41	0.349	0.523	0.528	0.333	0.405	0.415	0.303	0.056	0.277	0.212	0.171
nlp_123	42	0.237	0.357	0.432	0.369	0.310	0.304	0.034	0.118	0.040	0.075	0.231
CKingCoder	43	0.133	0.284	0.385	0.299	0.298	0.267	-0.123	0.048	0.212	0.145	0.100
CSECU_DSG	44	0.014	-0.003	-0.005	0.028	0.015	0.020	0.018	0.048	0.005	0.025	0.055
uaic_mt_2023	45	0.004	-0.066	-0.006	0.042	0.102	0.010	0.060	-0.082	0.005	0.008	-0.010

Table 5: Final leaderboard of SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis. 18 teams are able to beat the XLM-T baseline and 37 teams beat the mBERT baseline.

5.3 In-domain LM fine-tuning

Existing studies suggest that domain-focused language model fine-tuning could help to improve downstream task performances (Gururangan et al., 2020). As in-domain LM fine-tuning requires extensive computational resources, only 2 out of the 22 teams conduct in-domain LM fine-tuning. OPI (Dadas, 2023) fine-tuned 155M tweets over XLM-R with mask-language-modeling and achieved the best performance on 6 languages. Moreover, for participants using XLM-T, the XLM model fine-tuned on twitter data generally achieved better performance than participants using other base PLMs, suggesting that in-domain LM fine-tuning is still a simple yet effective approach to improve intimacy prediction in a multilingual setting.

5.4 Data augmentation

Given the multilingual nature of this task, data augmentation is actively used by participants. Fourteen out of the 22 participating teams used at least some form of data augmentation method for their final submissions.

Translation As part of the task is to predict in-

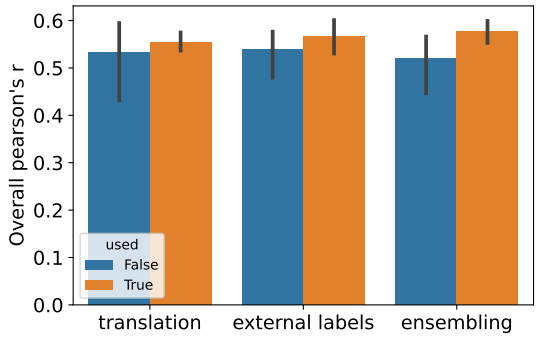

Figure 5: Teams using translation, external labels, and ensembling methods tend to perform better overall. However, due to the large variances, whether adopting these methods is significantly better remains unclear.

timacy in unseen languages, translation is widely used by participants as a way of data augmentation. Participants report performance improvements in ablation studies. Some participants (e.g., CHAPAT (Chavan and Patwardhan, 2023)) also tried to translate unseen languages back to English but only found marginal performance gain.

Externally labeled dataset Three out of 22 participating teams leveraged the question intimacy dataset (Pei and Jurgens, 2020) which contains 2397 questions from Reddit, Twitter, books, and movies annotated with intimacy scores from -1 to 1. Ablation studies were not reported in the system paper, therefore, the effectiveness of adding the question intimacy data remains unclear. UZH_CLYP (Michail et al., 2023) uses ChatGPT to generate labels for each language and shows performance improvement, suggesting that this approach may be helpful. OPI (Dadas, 2023) also leveraged pseudo-labelling (Lee et al., 2013) and achieved the best performance on 6 languages.

Rebalancing Due to the skewed distribution of the labeled data, participants also explored methods to either directly rebalance the training set or leverage learning algorithms to dynamically weight different samples. LAZYBOB (Yuan and Chen, 2023) used weighted random sampler (He and Garcia, 2009) to allow the model to focus more on interesting examples. SEA_AND_WINE (Chen et al., 2023b) designed an ad-hoc sampling strategy during data augmentation to balance the positive and negative samples. WADER (Suri et al., 2023) designed a weakly-labeling framework that includes distribution-based sampling and difference-based sampling methods to attain a more balanced training set.

Others Participants also explored other methods of data augmentation. HULAT (Segura-Bedmar, 2023) explored synonym replacement using EDA and ZBL2W (Zhang et al., 2023) experimented with word-level substitution. However, it is unclear whether these methods are helpful, due to the lack of ablation studies.

5.5 Ensemble

Many participants chose to combine multiple models' outputs as the final scores. The participants mostly explored two ways of ensembling: (1) fine-tune the same multilingual pre-trained model with different seeds and ensemble them for the final submission or (2) use separate models for different languages. Participants using ensemble methods generally performed better than others. The top 4 submissions (LAZYBOB (Yuan and Chen, 2023), UZH_CLYP (Michail et al., 2023), OPI (Dadas, 2023) and TMN (GLAZKOVA, 2023)) all used ensembles as part of their solution.

5.6 Other methods

While most of the systems generally focus on the methodologies above, some also explored other methods to improve their model performance. SEA_AND_WINE (Chen et al., 2023b), ZHEGU (He and Zhang, 2023), and ODA_SRIB (Kumar et al., 2023) explored different loss functions other than the standard MSE loss. All report performance gain with their specially designed loss function. Despite the regression nature of this task, ODA_SRIB explores binary classification problems with soft labels and optimizes the model with a BCE loss. Through ablation studies, ODA_SRIB found that a binary classification setting with BCE loss attained better performance than a regression with MSE loss.

Besides modifying the loss function, adversarial methods are also adopted by some participants: ODA_SRIB (Kumar et al., 2023) and LAZYBOB (Yuan and Chen, 2023) found that Adversarial Weight Perturbation (AWP) improves the final prediction performance.

6 Discussion

Intimacy is an important dimension of human relationships and language. Predicting intimacy across languages is a difficult task, because language is embedded in different cultures with different perceptions of intimacy. To address this issue, we built

MINT, the multilingual intimacy dataset. Forty-five participating teams have submitted their solutions to our multilingual tweet intimacy prediction task. Overall the participants are able to achieve relatively high performance (Pearson's rs > 0.7) on languages that are in the training set. While the performance on the unseen languages is relatively lower, many participants are able to improve their performance over the already-strong XLM-T baseline.

We observe a clear trend that most of the participants focused on the data augmentation and ensemble methods, with less effort spent on traditional feature engineering. Moreover, despite the fact that in-domain LM fine-tuning has been found to be helpful for downstream tasks (Gururangan et al., 2020), only a very limited number of participants considered in-domain LM fine-tuning as part of their solution. This is potentially due to the computational resources required for large-scale fine-tuning.

Despite recent advances in prompting, only one participant explored soft prompts, suggesting that the regression task remains challenging for prompting large language models. Our dataset provides a valuable multilingual resource for studying prompting regression tasks in the context of understanding social information. We encourage future researchers to explore this direction.

As a first step towards understanding textual intimacy in language, in this task, we only released aggregated labels from all the annotators. However, the perception of intimacy may vary with factors like gender, age, and culture. We are planning to release detailed background information of all the annotators. This will allow a potential follow task to predict intimacy ratings that incorporate annotators' background information.

7 Conclusion

In this paper, we present *SemEval 2023 Task 9: Multilingual Tweet Intimacy Analysis* along with MINT, the first multilingual textual intimacy analysis containing 13,372 tweets in ten languages (English, French, Spanish, Italian, Portuguese, Korean, Dutch, Chinese, Hindi, and Arabic). We benchmarked a series of multilingual pre-trained language models. Our task attracted participation from 45 teams and 18 teams are able to beat the already strong XLM-T baseline. Our overview indicates that data augmentation, external labels, and ensemble methods are commonly used by the participants, and led to good performance. Overall, participants generally achieved good performance on training languages with an average Pearson's r above 0.7, showing that the current pretrained language models, once fine-tuned, are able to accurately predict intimacy in various languages. However, the overall performance remains low on unseen languages, suggesting that zero-shot intimacy analysis remains a challenging task. Further research is needed to better analyze textual intimacy across languages.

Acknowledgment

Our multilingual tweet intimacy annotation has been approved by the IRB office of the University of Michigan (HUM00214259). This work was supported in part by the National Science Foundation under Grant No. IIS-2143529. We thank Jane Im, Minje Choi, Anubha Singh, Isra Salah, Abdulrahman Alhourani and Maria Casamor Vidal for help translating the task guidelines. We thank all the annotators on the Prolific platform. We thank all the participating teams for their contributions to this task.

References

Giorgos Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. 2023. Duth at semeval-2023 task 9: An ensemble approach for twitter intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1259–1264, Toronto, Canada. Association for Computational Linguistics.

Jin Yeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 60–64. Association for Computational Linguistics.

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996.

Prem Balasubramanian, Harish Kumar B, Naveen D, and Aarthi Gopinath. 2023. Ckingcoder at semeval-2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2043–2047, Toronto, Canada. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual

language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Abdessamad Benlahbib and Achraf Boumhidi. 2023. Nlp-lisac at semeval-2023 task 12: Sentiment analysis for tweets expressed in african languages via transformer-based models. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 199–204, Toronto, Canada. Association for Computational Linguistics.

Nimet Beyza Bozdag, Tugay Bilgis, and Steven Bethard. 2023. Arizonans at semeval-2023 task 9: Multilingual tweet intimacy analysis with xlm-t. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1689–1692, Toronto, Canada. Association for Computational Linguistics.

Tanmay Chavan and Ved Patwardhan. 2023. Chapat at semeval-2023 task 9: Text intimacy analysis using ensembles of multilingual transformers. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1337–1343, Toronto, Canada. Association for Computational Linguistics.

Yu Chen, You Zhang, Jin Wang, and Xuejie Zhang. 2023a. Ynu-hpcc at semeval-2023 task 6: Legal-bert based hierarchical bilstm with crf for rhetorical roles prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2109–2115, Toronto, Canada. Association for Computational Linguistics.

Yuxi Chen, Yu Chang, Yanqing Tao, and Yanru Zhang. 2023b. Sea_and_wine at semeval-2023 task 9: A regression model with data augmentation for multilingual intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 77–82, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Paul C Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin*, 79(2):73.

Slawomir Dadas. 2023. Opi at semeval 2023 task 9: A simple but effective approach to multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 150–154, Toronto, Canada. Association for Computational Linguistics.

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

José Antonio García-Díaz, Camilo Caparros-Laiz, Ángela Almela, Gema Alcaráz-Marbol, María José Marín-Pérez, and Rafael Valencia-García. 2023. Umuteam at semeval-2023 task 12: Ensemble learning of llms applied to sentiment analysis for low-resource african languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 287–294, Toronto, Canada. Association for Computational Linguistics.

Anna Glazkova. 2023. Tmn at semeval-2023 task 9: Multilingual tweet intimacy detection using xlm-t, google translate, and ensemble learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1350–1356, Toronto, Canada. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Peter Hartley. 2002. *Interpersonal communication*. Routledge.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Pan He and Yanru Zhang. 2023. Zhegu at semeval-2023 task 9: Exponential penalty mean squared loss for multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 320–325, Toronto, Canada. Association for Computational Linguistics.

Robert A Hinde. 1981. The bases of a science of interpersonal relationships. *Personal relationships*, 1:1–22.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Robert L. Johnson and Jim Penny. 2022. Split-half reliability. *The SAGE Encyclopedia of Research Design*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Sidney M Jourard and Paul Lasakow. 1958. Some factors in self-disclosure. *The Journal of Abnormal and Social Psychology*, 56(1):91.

Priyanshu Kumar, Amit Kumar, Jiban Prakash, Prabhat Lamba, and Irfan Abdul. 2023. Oda_srib at semeval-2023 task 9: A multimodal approach for improved intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1709–1713, Toronto, Canada. Association for Computational Linguistics.

Jelena Lazi and Sanja Vujnovi. 2023. The jelenasteam approach to multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 646–651, Toronto, Canada. Association for Computational Linguistics.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nirmal Manoj, Sagar Joshi, Ankita Maity, and Vasudeva Varma. 2023. irel at semeval-2023 task 10: Multi-level training for explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1724–1729, Toronto, Canada. Association for Computational Linguistics.

Abraham Harold Maslow. 1981. *Motivation and personality*. Prabhat Prakashan.

Andrianos Michail, Stefanos Konstantinou, and Simon Clematide. 2023. Uzh_clyp at semeval-2023 task 9: Head-first fine-tuning and chatgpt data generation for cross-lingual learning in tweet intimacy prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1050–1058, Toronto, Canada. Association for Computational Linguistics.

Roger E Millsap and Alberto Maydeu-Olivares. 2009. *The SAGE handbook of quantitative methods in psychology*. Sage Publications.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Jackson Sargent, Apostolos Dedeloudis, and David Jurgens. 2022. Potato: The portable text annotation tool. *arXiv preprint arXiv:2212.08620*.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326.

Abel Pichardo Estevez, Jacinto Mata, Victoria Pachón Álvarez, and Nordin El Balima Cordero. 2023. I2c-huelva at semeval-2023 task 9: Analysis of intimacy in multilingual tweets using resampling methods and transformers. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 773–777, Toronto, Canada. Association for Computational Linguistics.

Karen Jean Prager. 1995. *The psychology of intimacy*. Guilford Press.

Mohammadmostafa Rostamkhani, Ghazal Zamaninejad, and Sauleh Eetemadi. 2023. Rozam at semeval 2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2063–2066, Toronto, Canada. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Isabel Segura-Bedmar. 2023. Hulat at semeval-2023 task 9: Data augmentation for pre-trained transformers applied to multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 177–183, Toronto, Canada. Association for Computational Linguistics.

William E Snell, Rowland S Miller, and Sharyn S Belk. 1988. Development of the emotional self-disclosure scale. *Sex Roles*, 18:59–73.

Harry Stack Sullivan. 2013. *The interpersonal theory of psychiatry*. Routledge.

Manan Suri, Aaryak Garg, Divya Chaudhary, Ian Gorton, and Bijendra Kumar. 2023. Wader at semeval-2023 task 9: A weak-labelling framework for data augmentation in text regression tasks. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1979–1986, Toronto, Canada. Association for Computational Linguistics.

Rob van der Goot. 2023. Machamp at semeval-2023 tasks 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12: On the effectiveness of intermediate training on an uncurated collection of datasets. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 232–247, Toronto, Canada. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Mengfei Yuan and Cheng Chen. 2023. Lazybob at semeval-2023 task 9: Quantifying intimacy of multilingual tweets with multi-task learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 949–955, Toronto, Canada. Association for Computational Linguistics.

Hao Zhang, Youlin Wu, Junyu Lu, Zewen Bai, Jiangming Wu, Hongfei LIN, and Shaowu Zhang. 2023. Zbl2w at semeval-2023 task 9: A multilingual fine-tuning model with data augmentation for tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 785–790, Toronto, Canada. Association for Computational Linguistics.

Qinyuan Zheng. 2023. Wku_nlp at semeval-2023 task 9: Translation augmented multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 1558–1563, Toronto, Canada. Association for Computational Linguistics.