Modeling the Differential Prevalence of Online Supportive Interactions in Private Instant Messages of Adolescents

Ondrej Sotolar and **Michał Tkaczyk** and **Jaromir Plhak** and **David Smahel**
IRTIS, Masaryk University, Czechia
xsotolar@fi.muni.cz

Abstract

This paper focuses on modeling gender-based and pair-or-group disparities in online supportive interactions among adolescents. To address the limitations of conventional social science methods in handling large datasets, this research employs language models to detect supportive interactions based on the Social Support Behavioral Code and to model their distribution. The study conceptualizes detection as a classification task, constructs a new dataset, and trains predictive models. The novel dataset comprises 196,772 utterances from 2165 users collected from Instant Messenger apps. The results show that the predictions of language models can be used to effectively model the distribution of supportive interactions in private online dialogues. As a result, this study provides new computational evidence that supports the theory that supportive interactions are more prevalent in online female-to-female conversations. The findings advance our understanding of supportive interactions in adolescent communication and present methods to automate the analysis of large datasets, opening new research avenues in computational social science.

1 Introduction

For youth nowadays, social media, encompassing social network sites (SNS) and instant messaging applications (IM), have become significant avenues for seeking, offering, and receiving support, as well as for fostering a sense of belonging and emotional assistance (Wang et al., 2019). Because adolescents spend 2-3 hours per day using various social media platforms to communicate, primarily with their peers (Valkenburg et al., 2022), those platforms became crucial sites where adolescents establish their personal networks and social ties (Blahošová et al., 2023). At the same time, providing social support is an essential function of a person's social networks (Lu and Hampton, 2017). Being online, adolescents may easily engage in supportive interactions (SI) that serve as sources of perceived social support (Lin et al., 1979; Oh et al., 2014). Approximately two-thirds (68%) of American teenagers aged 13-17 have reported that social media make them feel as if they have people who will support them during challenging times (Center, 2022).

In recent years, the crucial role of social media in providing social support has gained recognition in academic research. Social scientists have invested considerable effort to understand how people seek and provide support online and how online social support affects individuals involved in supportive interactions (e.g., Liu et al., 2018; Utz and Breuer, 2017; Chang, 2009). In these works, social scientists rely mainly on self-reports when analyzing online supportive interactions and their effects. However, self-reported measurement of behavior, which is a conventional social science method (Gottlieb and Bergen, 2010), is characterized by low accuracy, low validity, and bias, that is, the tendency to over- or under-reporting of measured behaviour (Kormos and Gifford, 2014; Parry et al., 2021).

The observational measurement, which provides much more accurate insights into the communicative behavior of social media users, requires labor-intensive methods, such as manual content analysis (Krippendorff, 2018), which limits the size of the datasets that can be analyzed (e.g., Cheung et al., 2017) and makes the analysis of large datasets sourced from social media (Lewis et al., 2013) infeasible. Additionally, the conventional methods cannot be utilized in real-time assessment and intervention procedures such as ecological momentary interventions aimed at psychological health (Heron and Smyth, 2010). Therefore, the field would benefit from utilizing state-of-the-art AI, specifically language models, that could detect supportive interactions automatically.

The current study focuses on using such models as scientific tools. We pose the question of how

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 6223–6241

Label	Description
Informational Support	provide useful knowledge and information, feedback, or experience
Emotional Support	express intimacy, caring, liking, empathy, willingness to listen
Social Companionship	convey a sense of belonging, inclusivity, spending time together, recreational activities, invitation for participation
Appraisal	express acceptance, respect, validation, esteem, approval
Instrumental Support	offer practical help or resources, assistance in getting tasks done

Table 1: Overview and short definitions of labels based on the SSBC categories from Cutrona and Suhr (1992).

reliably we can detect supportive interactions in online conversations. Then, we formulate a practical experiment showcasing the usage of our models to automate the analysis of large datasets. We seek to verify hypotheses proposed in the existing literature that in dialogues, the prevalence of supportive interactions is different between groups of participants with different characteristics, such as gender makeup (Tifferet, 2020; Zhou et al., 2017; Andalibi et al., 2017; Reevy and Maslach, 2001) or communication in pair and a group (Bambina, 2007). Validating such methods against theories already supported by statistical evidence shows that modeling distributions using the predictions of language models is a viable alternative when datasets are too large to employ manual methods.

To answer the research questions, we collected data from volunteer adolescents (ages 13-17, $M_{age} = 15.86$, 36% women) from Instant Messenger apps Messenger and WhatsApp. IM is a type of online communication that allows for the synchronous exchange of text and multimedia between two or more people (Huang and Leung, 2009). Such private online communication has been understudied so far (e.g., Huh-Yoo et al., 2023a; Ali et al., 2023; Underwood et al., 2012). IM conversations are suitable for studying supportive interactions because communication between users involved in close relationships is the primary source of perceived social support, which typically occurs in private communication (Cutrona and Suhr, 1992). We annotated 196,772 utterances in Czech authored by 2165 users, creating a new dataset annotated with five different categories of supportive interactions based on the Social Support Behavioral Code (SSBC, Cutrona and Suhr, 1992): *Informational Support*, *Emotional support*, *Social Companionship*, *Appraisal*, and *Instrumental support* (see Table 1). We cast the detection problem as classification and compare models of different sizes and architectures to show the detection feasibility, including comparing our approach to the best approach from previous work. Finally, we analyze and compare the predicted and ground truth distributions of supportive interactions and provide new computational evidence for the investigated theories using statistical tests. We publish both our code (Github-Repository, 2024) and the trained models (HuggingFace-Hub, 2024) to enable replication of our results.

2 Related Work

A broad view of the application of AI and machine learning methods for detecting phenomena related to mental health (Le Glaz et al., 2021; Rooksby et al., 2019) in online and social network data (Zhao et al., 2022; Mendu et al., 2020) shows a vast selection of literature in multiple fields (Chancellor and De Choudhury, 2020; Thieme et al., 2020), for example for sentiment or opinions analysis (Neethu and Rajasree, 2013; Jin et al., 2009; Sidorov et al., 2013) or answering questions about social trends (Das et al., 2015; Chen et al., 2021). However, considering the unique domain of private communication through instant messaging, the selection of direct predecessors to our work narrows significantly, presumably because collecting such data is difficult.

IM conversations of adolescents have been explored for different research goals, such as to determine if participation allows practicing social skills or forming offline relationships (Koutamanis et al., 2013), exploring one's identity, and finding information (Valkenburg and Peter, 2011) as well detecting a variety of online risks, such as cyber-aggression (Álvarez García et al., 2018) or online solicitation (Valkenburg and Peter, 2011). The researched topics mainly cover mental health, often aimed at detection and diagnosis (Shatte et al., 2019). Importantly, there is a profound lack of re-

search on positive influence factors, as the research focuses more on negative aspects, such as online risks. Many challenges in this domain prevail, such as conceptualizing mental health, the diversity of mental health problems, the sparsity of data, and the multi-modality and multilingualism of corpora from social networks (Rahman et al., 2020). Studies mostly use public datasets containing data from platforms such as Twitter (Al-garadi et al., 2016) and YouTube (Dadvar et al., 2013). Due to the significant domain shift compared to IM conversations, methods and results from these works are not directly applicable or comparable to IM data (Rosa et al., 2019).

Our selection of references is focused on supportive interactions and not on other conceptually unrelated types of communicative behavior because different psychological phenomena differ in terms of their linguistic features at the interactional or ideational levels of language.

2.1 Social Support Online

Cobb (1976) defined social support as "information leading the subject to believe that he or she is cared for and loved, that he/she is esteemed and valued, and he/she belongs to a network of communication and mutual obligation". This early definition has already encompassed some of the functional dimensions of the constructs that were identified in subsequent research (e.g., Cutrona and Suhr, 1992; Wills and Shinar, 2000). Differentiating between different types of support is important because an individual's preferences for support may be different, and only when the received social support matches one's needs does it result in improvement, psychological adjustment, and ability to cope with distressing events (Andalibi et al., 2017). In this study, we categorized supportive interaction according to the Social Support Behavioral Code developed in Cutrona and Suhr (1992), which is seminal and the most nuanced categorization schema. It was developed to assess 23 support-intended communication behaviors that fall into five categories. *Informational Support* is support through providing helpful information, such as giving advice or providing feedback. *Emotional support* lies in communicating love, care, or empathy. *Companionship or social network support* is provided by communicating belonging to a group with similar interests or concerns. *Esteem support* lies in communication, respect, understanding, and confidence in one's abilities. Finally, *tangible help aid* is performed through providing or offering goods and services.

In the literature, a clear differentiation has been established between two forms of social support: perceived support and enacted support. Perceived social support is the outcome of supportive communication. It can be understood as subjective perception – that is, a person's belief that they experience the feeling of being supported by his or her social ties (Lin et al., 1979). Enacted support comprises actual behaviors, and it can be defined as the exchanges of resources or aids between individuals through interpersonal ties (Oh et al., 2014). More recent literature discriminates between offline and online social support (Oh et al., 2014; Tifferet, 2020). The latter concept refers to "social support received via any means of online communication" (Utz and Breuer, 2017), which is the focus of this study.

As youth spend more time online, supportive communication shifts to online platforms. In addition, several online environment features make social media a convenient space for seeking and obtaining support. Factors such as the absence of nonverbal cues facilitate more intimate disclosures (Tidwell and Walther, 2002). At the same time, the social distance between individuals is greater, and interaction management is easier as compared to face-to-face interactions (Joseph B. Walther and Shawn, 2002). Consequently, the Internet and social media became important sites where people seek and receive social support (Wang et al., 2019). Several studies found positive associations between the number of SNS friends, frequency of social media use, online supportive interactions, affect, perceived social support, and sense of community satisfaction (Lu and Hampton, 2017; Oh et al., 2014). For people who use social media, supportive interactions with other individuals that take place online are the primary source of perceived support from others (Oh et al., 2014).

Crucially, the support-related opportunities provided by online media might vary among individuals and contexts. In this study, we explore gender-related differences and differences stemming from one-to-one versus group character of conversation. Differences were found between girls and boys in seeking and providing social support online and offline. While prior findings are equivocal, results of the meta-analysis showed that females on SNS give ($d = 0.36$) and receive ($d = 0.14$) greater social support as compared to males (Tifferet, 2020). It was also found that women are more likely to offer

social support to their same-sex friends than men to their same-sex friends (Zhou et al., 2017). One explanation of this pattern is gender roles assuming that women are socialized to believe that the norm is to be emotional, talk about their problems, and seek help (Andalibi et al., 2017). Gender differences also manifest in the type of support being sought or provided. Femininity (for both sexes) was found to be associated with seeking and receiving emotional support, and masculinity with receiving tangible support (Reevy and Maslach, 2001). Concerning differences in support provision between one-to-one and group conversations, it was found that different support categories are associated with factors like different levels of intimacy or symmetry between group participants. At the same time, those factors are partly related to the number of people participating in interaction (Bambina, 2007).

In conclusion, social science literature considers mostly self-reported evidence, without considering real-world setting, and with low ecological validity (e.g., Huh-Yoo et al., 2023b). Importantly, composite measures of time spent with social media or IM, do not differentiate specific practices on the platform. Lu and Hampton (2017) shows that different ways in which people interact with a social media platform will be related to different outcomes also in terms of social support; therefore, **directly measuring** supportive interaction that takes place in native communication context would allow studying prevalence, predictors, and effects of different types of youth peer support that happens in social media with high ecological validity.

2.2 Detecting Psychological Phenomena in Private Messaging

Studies on informational support (Williamson and O'Hara, 2017; Feng and Magen, 2016; High and Buehler, 2019) and social companionship (Treré, 2015) use metadata such as the number of messages, number of calls, sent/received ratio, time spent, social network activity, and others. Appraisal has not been the target of predictive modeling; however, it is discussed in the context of social networks in Feng and Hyun (2012). For Emotional Support, some feature-based models successfully use features such as emoticon count and type, as in Xu et al. (2007). Instrumental Support in IM conversations has been studied (not with predictive models) in Xie (2008).

Considering the data domain of IM, Underwood et al. (2012) pioneered the automated analysis of private messages by collecting and analyzing the BlackBerry SMS corpus. With this and several subsequent works (Skierkowski and Wood, 2012; Underwood et al., 2015; Brinkley et al., 2017), the authors have built successfully trained predictive models using feature-based machine learning methods. Subsequently, Nobles et al. (2018) have also collected a dataset of SMS messages and introduced a classifier based on the combination of feature-based modeling and multi-layer perceptron neural network architecture. Recently, authors have started analyzing data from social network sites because they also enable private communication between users in addition to public-facing communication. Ali et al. (2023) have collected data from Instagram direct messages and compared a range of feature-based machine learning classifiers to a convolutional neural network from Kim (2014) which showed the best results. The dataset has been further analyzed in subsequent work (Razi et al., 2023; Huh-Yoo et al., 2023a). Plhák et al. (2023b) have analyzed Messenger and WhatsApp data of adolescents using transformer-based models. They have shown that leveraging the context of the classified utterances helps predict risky behavior. Sotolář et al. (2024) explore the possibility of detecting common positive and negative influence factors that impact adolescents' well-being in instant messenger communication and they show that leveraging the similarities between concepts can improve the success of the detection.

Concerning methods, NLP has generally moved away from feature-based modeling, which has inherent shortcomings, as shown in Bahdanau et al. (2015); Vaswani et al. (2017); Radford et al. (2018); Yang et al. (2019). However, the penetration of representation-based models to other fields might be slow because even recent works, such as Nobles et al. (2018); Razi et al. (2023), explore feature-based statistical machine learning classifiers.

Datasets in the domain of private dialogues are usually not published for privacy concerns. We have inquired the authors of Underwood et al. (2012), Nobles et al. (2018), and Razi et al. (2023) for data, but we were denied access based on privacy concerns. Therefore, we cannot directly redo the experiments from the referenced works. However, we compare our methods to the best method from Razi et al. (2023) on our data.

		Supportive Interaction					
		Infor. Support	Emotional Support	Social Compan.	Appraisal	Instrum. Support	All incl. 'none'
Grammatical gender of author names	F	53.2	55.6	45.6	54.6	53.8	46.8
	M	46.8	44.4	54.4	45.4	46.2	53.2

Table 2: Distribution of supportive interactions in conversations across demographic factors (% of total).

3 Dataset

We collected a dataset of online private conversations that the participants of our research led with their counterparts, which were exported from the Messenger and WhatsApp applications. Each of the participants (13-17 years old) conversed with multiple counterparts; the final dataset contains data authored by 2165 different users, 90,422 individual conversations with 1,260,492 utterances in total. Research such as Benotsch et al. (2013); Valkenburg and Peter (2011) shows that adolescents overwhelmingly communicate with peers, which allows us to assume that most of the dataset is composed of such conversations. 87.25% of utterances are in Czech, 8,08% in Slovak (Czech and Slovak are mutually intelligible sister languages), and 4.67% in English. 53.2% of users were male and 46.79% female (refer to Table 2), which we determined by automatic morphological analysis of the users' names using tools from Straková et al. (2014). We provide further demographics and authorship distribution statistics in Appendix A.

Annotated Dataset We annotated the data using five labels based on the five categories of supportive interactions as defined in the SSBC as shown in Section 1, Table 1. We added the 'none' label for utterances without any instances present. Each example was annotated by two raters and the inter-annotator agreement (IAA) was measured with Cohen's κ. To create the gold-standard version of the dataset, where all examples have at least two-vote confidence, the label conflicts were resolved by an additional round of annotation. The statistics of the resulting dataset are shown in Table 3. We provide further details on the annotation process and additional dataset statistics in Appendix A.

4 Methods

4.1 Data Preprocessing and Preparation

Plhák et al. (2023b) have shown that classifying utterances along with context is beneficial for detecting phenomena in dialogues. Therefore, our method defines the dataset examples as *local context windows* within the dialogues. The context helps the models capture local dependencies; e.g., a winking emoticon following an utterance may suggest its meaning was joking or sarcastic.

We do not use whole dialogues as examples mainly because it diverges conceptually from our research questions – we aim to detect local instances of supportive interactions. Furthermore, using such examples requires different evaluation metrics, such as earliness of detection (Vogt et al., 2021). Nevertheless, some examples contain whole dialogues because some can be shorter than the context windows. The ratio of such examples depends on the size of the context windows. For the utterance length statistics (Figure 1) and the distribution of possible example-length composition (Figure 2), refer to Appendix A.

While the data is annotated with five subcategories of supportive interactions, training the model to discriminate the individual subcategories would not be useful for the goal of the current study – that is, predicting the distribution of supportive interactions in a dataset (see Section 4.4). Furthermore, Sotolář et al. (2024) have shown that the individual categories of supportive interactions are closer to each other in the model's latent space than examples without supportive interactions; therefore, aggregating the fine-grained labels into coarser labels improves the detection success. We utilize this finding and binarize the problem by aggregating the subcategory labels into two binary *yes/no* labels denoting the occurrence of supportive interactions.

To construct the examples, we use a sliding window starting from the last utterance to capture the local context. The window size is guided by the maximum input width of the model, and to delimit the window size precisely, we use a soft character limit – breaking on whole utterances. For the context window construction diagram, refer to Fig-

Supportive Interaction	Number of labels	P(%)	Cohen's κ	Number of Gold-Standard labels	P(%)
Informational support	9967	5.07	.685	7322	3.72
Emotional Support	9669	4.92	.639	10526	5.35
Social Companionship	3331	1.7	.604	5766	2.93
Appraisal	2338	1.19	.65	2524	1.28
Instrumental Support	5317	2.7	.599	3640	1.85
None	166150	84.42	–	168096	84.87
All Categories	30622	15.58	–	28676	15.13
All Labeled Utterances	196772	100	–	196772	100

Table 3: Utterance annotation counts in the corpus and the inter-annotator agreement. The left part refers to the labels produced by two annotators before the supervisor resolves conflicts, and the right part refers to the labels after the supervisor provides the deciding third vote.

ure 3 in Appendix A. Because the context window contains a sequence of labels, we produce the final label of an example by aggregating the labels: if the sequence of labels contains any that positively labels one of the subcategories, the whole example is assigned the positive label.

4.2 Evaluation Metrics, Reliability, and Class Imbalance

For training, model selection, and validation, we split the dataset with the target ratio of 80:5:15 respectively (N = 196,772, n_{train} = 157,429, n_{dev} = 9835, n_{test} = 29,508).

To evaluate the prediction, we use the bootstrapping method from Efron (1979) to estimate confidence intervals. It involves repeatedly drawing subsamples with replacement from the validation sample, creating many simulated test subsamples, and measuring the predictive performance with the F_1 measure on the subsamples. By analyzing the distribution of the prediction performance, we derive a 95% confidence interval. To check the statistical significance when comparing two classifiers, we derive the p-values using the permutation test (Dror et al., 2018).

For classification, the imbalance of class distribution (refer to Appendix A, Table 3) needs to be addressed. We evaluated three approaches: weighting the loss function, augmenting by adding paraphrases, and simple oversampling. The latter two showed better results than weighting, but the augmentation did not significantly improve over oversampling. Therefore, we used the simpler oversampling approach for the training data while the validation sample retained its class distribution.

4.3 Classification Models

We use the cannonical classifier architecture: a classification head with the *softmax* function on top of a pre-trained transformer model. We selected fine-tuning as the model adaptation strategy based on the number of labeled examples (refer to Table 3). We fine-tune the models using the *cross-entropy* loss function. We trained with early-stopping until the loss function saturation. For the detailed training setup, see Appendix B.

We evaluate several notable pretrained models to determine the effects of model size, the pre-trained weights, and architecture. For pre-trained bases, we compare the parameter-efficient ELECTRA model from Kocián et al. (2022) (Small-E-Czech), the RoBERTa model from Straka et al. (2021) (RobeCzech), and the much larger GPT-3 (Brown et al., 2020). We also use the multilingual model XLM-RoBERTa (Conneau et al., 2020), which allows the models to be used on data in any of the model's pre-training languages (such as English) using cross-lingual transfer, and we expand on it in Section 4.5.

4.4 Modeling Gender-based and Pair-or-Group Differences in the Prevalence of Supportive Interactions

The prior social science research showed gender-based differences in the prevalence of supportive interactions. It also showed that the prevalence varies between one-to-one and group communication. Our method, which provides computational evidence for such claims, can be broken down into three steps: We detect the characteristics of the par-

ticipants of a conversation, and within it, we detect the supportive interactions using the classification models presented in this study. Finally, we compare the distributions of the detected supportive interactions to the distribution in the ground truth from the annotated data.

We define two samples of conversations characterized by the following:

1. Gender-makeup of participants: *all-female*, *all-male*, *mixed*.

2. Number of participants: *1-on-1 conversations*, *group chats*.

We performed morphological analysis to draw the subsample from the test sample based on the gender makeup (for detailed statistics, refer to Appendix A.2). The subsample for the number of participants was drawn by counting the distinct authors. We evaluate this experiment in Section 5.2 by using the Z-test on the predicted ratios of supportive interaction labels in the subsamples. We use the Z-test to compare the proportions of different labels between two samples because it is designed for testing differences in proportions with large sample sizes (PennState, 2024). The test checks if the proportion of a categorical outcome significantly differs between groups, assuming the sample sizes are large enough for the sampling distribution to approximate normality.

4.5 Using the Models in Other Languages

The majority language of the data is Czech (see Section 3), but multilingual models like XLM-RoBERTa (Conneau et al., 2020) can be fine-tuned on data in many languages. We use two methods to achieve this: first, we translate our entire dataset using the DeepL machine translation service (Kutylowski) and train and test the models in the same way as with the original data, and second, we leverage the cross-lingual ability of the XLM-RoBERTa model, which promises to enable the usage of a model trained on data in one language to produce predictions also for data in other languages.

5 Results

This section presents the effectiveness of supportive interaction detection in private dialogues and the utility of the trained models for modeling gender-based and pair-or-group disparities.

5.1 Effectiveness of the Classification Models

The Effect of Model Size In Table 4, we compare models of different sizes and architectures. The four larger models performed similarly to each other – the difference was not statistically significant ($p > 0.05$), but all performed significantly better than the smallest model ($p < 1e-5$). We conclude that the models are limited by the quality and quantity of the annotated data (see the IAA in Table 3) rather than the model size. Therefore, in subsequent experiments, we use the smallest of the large multilingual models (XLM-RoBERTa-large).

Base Model	Par.	$F1 \pm CI$	AUC
Small-E-Czech	14M	80.15 ± 1.5	.737
RobeCzech-base	127M	86.78 ± 1.6	.782
XLM-R-L	561M	86.53 ± 1.6	.822
XLM-R-XL	3.5B	86.27 ± 1.6	.821
GPT-3	175B	86.34 ± 1.4	.811

Table 4: The comparison of classification results using models of different sizes and architectures measured with the $F1$ measure with estimated 95% confidence intervals and the Area Under Curve (AUC) metric.

Comparing our Best Model to Previous Work We compare our best model to the best method from Razi et al. (2023), the most recent related work. Our model significantly outperformed ($p < 1e-5$) the referenced model by a large margin.

Model	$F1 \pm CI$	AUC
CNN - best in Razi et al. (2023)	71.82 ± 1.8	.687
RobeCzech-base	86.78 ± 1.6	.782

Table 5: Comparison of classification results between our best model and the best model from Razi et al. (2023). The primary metric is $F1$ with estimated 95% confidence intervals and the Area Under Curve (AUC).

5.2 Modeling the Gender-based and Pair-or-Group Differences

We compared the distribution of occurrences of supportive interactions between the predictions made by our model and the annotated dataset, which holds the ground truth labels. The results are shown in Table 6.

In our dataset, there are statistically significant differences in the distribution of supportive interactions related to the detected grammatical gender of participants in the conversation. More specifically, the occurrence of supportive interactions is higher in conversations between-girls as compared to conversations between-boys and conversations between-participants-of-different-gender. We observed the same differences in the predictions of our model. With a series of Z-tests, we confirmed that there was no statistically significant difference between the predicted occurrence of supportive interactions and the ground truth for each of the gender groups.

Concerning the differences across conversations with different numbers of participants, there are no statistically significant differences in the occurrence of supportive interactions between 1-on-1 and group chats. We observed the same homogeneity with our model and confirmed the statistical significance of the result with a series of Z-tests.

5.3 Language Variations

In Table 7, we present the results of experiments with the machine-translated version of the dataset and the cross-lingual transfer method using the fine-tuned XLM-RoBERTa-large. The model performed the best while trained and evaluated on the original Czech data. The difference to the other models is significant ($p < 0.01$). The machine translation and cross-lingual transfer perform similarly ($p > 0.05$).

6 Discussion

Detected Increase of Supportive Interactions in the All-Female Conversations The results shown in Section 5.2 provide evidence that our models can reliably detect the differences in the frequency of occurrence of supportive interactions between the all-female, all-male, and mixed-gender dialogues. Despite the imperfect prediction (as indicated by F1 scores), the model was able to reliably detect patterns existing in the data and provide further evidence to what was shown in literature on social support: for example, Zhou et al. (2017) concludes that women are more likely to offer social support to their same-sex friends than are men to their same-sex friends and Tifferet (2020) showed that females on SNS give ($d = 0.36$) and receive ($d = 0.14$) greater social support as compared to males. Based on this, we conclude that we have evidence that our models can become valuable tools for social science research.

Applicability of the Models across Languages The results in Table 7 show that the highest performance was achieved with models trained on the data in the original language. The machine translation hurt the performance, and the difference is statistically significant but small. Therefore, we conclude that our models trained on the machine-translated version of our dataset are also usable for English data.

The experiments with cross-lingual transfer show exciting results: the transfer from Czech to English and vice versa shows results that are not statistically different from the machine-translated version. However, we conclude that our discriminative models, trained on the English dataset and all of the generative models, can be used in other languages covered by the pre-training dataset of the XLM-RoBERTa-large.

6.1 Exploring Correctly and Incorrectly Classified Examples

We use the gradient-based technique of Layer Integrated Gradients (LIG) from Atanasova et al. (2020); Robnik-Šikonja and Bohanec (2018) to get the estimated contribution weight of the input tokens. We use LIG to highlight tokens that contribute to the prediction proportionally to the intensity of its shade, green to the positive, and red to the negative classes (refer to Appendix C, Table 8). A pattern we observed from the LIG attributions is the expected model's bias towards the leftmost tokens, also shown in Catena et al. (2019); Wu et al. (2019).

Among the correctly classified (true-positives, TP), we have found mainly semantically sound and well-structured text as opposed to the false-negatives (FN), among which we found many examples where the text lacked any clear meaning and examples with ambiguous labels. We argue that this partially stems from the overall low quality of text in the language domain of IM messages, where the utterances are often just sentence fragments or short in length (see Appendix A, Figure 1), and partially from the dataset annotations which exhibit only a moderate IAA (see Appendix A, Table 3). The ambiguity of labels also affects (although to a lesser degree) the false-positive (FP) examples.

A unique pattern for the FP examples was the presence of unanswered questions – question-

Sample Characteristic	True Ratio of SI	Predicted ratio of SI	z-score	p-value	$F1 \pm CI$
all-female	$.6890^a$	$.6956^a$	-0.705	.484	81.87 ± 0.8
all-male	$.5660^b$	$.5716^b$	-0.544	.589	76.76 ± 1.1
mixed	$.5773^b$	$.5802^b$	-0.496	.617	75.28 ± 0.6
1-on-1	$.6027^a$	$.6042^a$	-0.281	.780	77.57 ± 0.4
group chat	$.5830^a$	$.5960^a$	-1.401	.162	75.39 ± 0.9

Table 6: The predictive analysis of characteristic subsamples of the test sample. The classification results with the fine-tuned XLM-RoBERTa-large model are measured with the $F1$ measure with estimated 95% confidence intervals. Cells sharing a letter in superscript (^{a} or ^{b}) are not significantly different by $\tilde{\chi}^2$ test of association, with p-values adjusted by false discovery rate method for multiple comparisons (see Benjamini and Hochberg, 1995).

lang(train)	lang(eval)	$F1 \pm CI$	AUC
cs	cs	86.53 ± 1.6	.822
en	en	84.58 ± 1.7	.787
cs	en	84.20 ± 1.8	.752
en	cs	84.93 ± 1.9	.807

Table 7: The comparison of classification results with the fine-tuned XLM-RoBERTa-large for the language variants of the dataset (Czech-original, English-machine-translated) and cross-lingual transfer measured with the $F1$ measure with estimated 95% confidence intervals and the Area Under Curve (AUC) metric.

answer pairs where the answer is curt and question-answer pairs that contain a negative answer or fail to provide the needed support. Many examples take the form of question-answer pairs, and the models did not discriminate well enough between the answered and unanswered ones compared to the annotators, who could discriminate well.

7 Conclusion

In this study, we presented and evaluated novel and newly applied methods to modeling supportive interactions in the rarely explored domain of Instant Messaging conversations among adolescents. Using trained language models, we could reliably model the distribution of supportive interactions in dialogues led by participants of different characteristics, specifically gender-based and pair-or-group disparities. We thus provided new computational evidence to validate theories concerning the differential prevalence of online supportive interactions. We confirmed the theorized gender and group-size differences: dialogues between girls showed an increase in supportive interactions as opposed to mixed-gender or all-boys, and we found that communication in pairs or groups had no significant impact on the prevalence. We have published our models for Czech and English, and with cross-lingual transfer, for 98 other languages. We also show the helpfulness of utilizing gradient-based explainability techniques for detecting common patterns in the error analysis for detecting supportive interactions. Overall, we found that modeling supportive interactions with language models is effective enough to become a valuable tool in computational social science research for automating the analysis of large datasets and providing computational evidence for theories.

8 Limitations

Annotation The annotation application's UI presented context windows of utterances to the annotators - a small part of the annotated dialogue. They could scroll to previous/following utterances, but for future work, it would be helpful to log this behavior because we could have used this data to determine the statistics on the annotator's decisions, such as the average context needed to decide on each example's label.

Reproducibility We adhere to strict ethical and legal considerations while working with adolescents' objective data (see Section 9). That limits our study, as we cannot publish our dataset. Although the dataset is sufficiently anonymized using a machine learning approach (Sotolář et al., 2021) and with an additional layer of safety provided by a non-disclosure statement by the annotators, it is impossible to release our dataset because the unacceptable risk of re-identifying the users by inference by a determined attacker acting in bad faith. Even the state-of-the-art anonymization approaches, such as Hu et al. (2023); Igamberdiev and Habernal (2023) have limitations that would prevent publishing the dataset. This is common in research

concerning human subjects, such as medicine or psychology. We have inquired the authors of Underwood et al. (2012), Nobles et al. (2018), and Razi et al. (2023) for data, but we were not granted access based on privacy concerns. However, we compare our methods to the best method from Razi et al. (2023) on our data (see Section 5.1).

Data Modality We omitted other modalities than text in data collection because, at the time no reliable anonymization method existed for modes such as images. However, we remain optimistic about the potential of future authors to find a solution or try to transfer the information from the data from other modes on the client side, thus bypassing the need for anonymization.

Robustness of Supportive Interaction Detection The distribution of utterance authors in the dataset is not optimal, as shown in Figure 8, which may lead to overfitting on the semantic, stylistic, or syntactic features of the users that contributed more data than others. It is a challenge we encourage future authors to consider and address using larger language models or larger datasets to enhance the robustness of the detection.

Applications Besides our methods being useful as scientific tools for big data and real-time analysis, future work might focus on integrating the presented methods into parental control applications. Such applications could inform parents about their children's supportive online interactions (or lack thereof) in addition to the more commonly studied negative influence factors, such as online risks.

9 Data Privacy and Ethical Considerations

Our research participants were volunteers, and they and their parents gave informed consent to participate. Due to the sensitivity of the data, we went above and beyond to adhere to the legal and ethical recommendations of the Research Ethics Committee of our institution, the GDPR (European Union, 2016), and the research protocol of the study (Elavsky et al., 2022) . All text messages were anonymized and access-protected on multiple levels. Before uploading to a private server through a custom desktop application, the data were anonymized with the method described in Sotolář et al. (2021); therefore, nobody had access to non-anonymized data. Metadata, such as the authors' names, were also anonymized. Therefore, the platforms that were data sources cannot be queried ex-post to retrieve user information. The chosen anonymization method is tuned for precision and was shown not to affect classification accuracy. Moreover, multimedia messages were replaced by appropriate tags (such as <photo>, <gif>, <audio>) to preserve anonymity and also continuity, such as reactions.

For uploading the data, we developed a custom digitally-signed desktop application that removed any multimedia content (see Section 8) and anonymized the conversations on the client side before uploading them over a secure channel to our secure and access-protected server. All who had access to the data signed a non-disclosure agreement. Access to the data was divided into three levels: administrators, researchers, and annotators. Only a randomized selection of samples was presented to the annotators through a web application for IM annotation (Plhák et al., 2023a) to prevent familiarization with authors' styles and discussed topics. The annotators passed intensive training in data confidentiality. If they found or suspected a rare case of re-identification or attribute disclosure, they reported the case, which was mitigated.

Acknowledgements

The data described/study is from the project „Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583“ which is co-financed by the European Union.

References

Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications. *Comput. Hum. Behav.*, 63(C):433–443.

Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–30.

Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1485–1500, Portland Oregon USA. ACM.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Antonina Bambina. 2007. *Online social support: the interplay of social networks and computer-mediated communication*. Cambria press.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Eric G Benotsch, Daniel J Snipes, Aaron M Martin, and Sheana S Bull. 2013. Sexting, substance use, and sexual risk behavior in young adults. *Journal of adolescent health*, 52(3):307–313.

Jana Blahošová, Michaela Lebedíková, Martin Tancoš, Jaromír Plhák, David Šmahel, Steriani Elavsky, Michal Tkaczyk, Ondřej Sotolář, et al. 2023. How are czech adolescents using their phones? analysis using objective smartphone data.

Dawn Y Brinkley, Robert A Ackerman, Samuel E Ehrenreich, and Marion K Underwood. 2017. Sending and receiving text messages with sexual content: Relations with early sexual activity and borderline personality features in late adolescence. *Computers in human behavior*, 70:119–130.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Matteo Catena, Ophir Frieder, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Nicola Tonellotto. 2019. Enhanced news retrieval: Passages lead the way! In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1269–1272.

Pew Research Center. 2022. Connection, Creativity and Drama: Teen Life on Social Media in 2022.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Hui-Jung Chang. 2009. Online supportive interactions: Using a network approach to examine communication patterns within a psychosis social support group in Taiwan. *Journal of the American Society for Information Science and Technology*, 60(7):1504–1517.

Yunsong Chen, Xiaogang Wu, Anning Hu, Guangye He, and Guodong Ju. 2021. Social prediction: a new research paradigm based on machine learning. *The Journal of Chinese Sociology*, 8:1–21.

Yee Tak Derek Cheung, Ching Han Helen Chan, Man Ping Wang, Ho Cheung William Li, and Taiching Lam. 2017. Online Social Support for the Prevention of Smoking Relapse: A Content Analysis of the WhatsApp and Facebook Social Groups. *Telemedicine and e-Health*, 23(6):507–516.

Sidney Cobb. 1976. Social support as a moderator of life stress. *Psychosomatic medicine*, 38(5):300–314.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Carolyn E. Cutrona and Julie A. Suhr. 1992. Controllability of Stressful Events and Satisfaction With Spouse Support Behaviors. *Communication Research*, 19(2):154–174.

Maral Dadvar, Rudolf Berend Trieschnigg, and Franciska MG de Jong. 2013. Expert knowledge for automatic detection of bullies in social networks. In *25th Benelux Conference on Artificial Intelligence, BNAIC 2013*, pages 57–64. Delft University of Technology.

Anubrata Das, Moumita Roy, Soumi Dutta, Saptarshi Ghosh, and Asit Kumar Das. 2015. Predicting trends in the twitter social network: a machine learning approach. In *Swarm, Evolutionary, and Memetic Computing: 5th International Conference, SEMCCO 2014, Bhubaneswar, India, December 18-20, 2014, Revised Selected Papers 5*, pages 570–581. Springer.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Steriani Elavsky, Jana Blahošová, Michaela Lebedíková, Michał Tkaczyk, Martin Tancos, Jaromír Plhák, Ondřej Sotolář, David Smahel, et al. 2022. Researching the links between smartphone behavior and adolescent well-being with the future-wp4 (modeling the future: understanding the impact of technology on

adolescent's well-being work package 4) project: protocol for an ecological momentary assessment study. *JMIR Research Protocols*, 11(3):e35984.

European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).

Bo Feng and Min Jung Hyun. 2012. The influence of friends' instant messenger status on individuals' coping and support-seeking. *Communication Studies*, 63(5):536–553.

Bo Feng and Eran Magen. 2016. Relationship closeness predicts unsolicited advice giving in supportive interactions. *Journal of Social and Personal Relationships*, 33(6):751–767.

Github-Repository. 2024. Code on github. https://github.com/csocsci/supportive-interactions.

Benjamin H. Gottlieb and Anne E. Bergen. 2010. Social support concepts and measures. *Journal of Psychosomatic Research*, 69(5):511–520.

Kristin E Heron and Joshua M Smyth. 2010. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *British journal of health psychology*, 15(1):1–39.

Andrew C High and Emily M Buehler. 2019. Receiving supportive communication from facebook friends: A model of social ties and supportive communication in social network sites. *Journal of Social and Personal Relationships*, 36(3):719–740.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2023. Differentially private natural language models: Recent advances and future directions. *arXiv preprint arXiv:2301.09112*.

Hanyun Huang and Louis Leung. 2009. Instant messaging addiction among teenagers in china: Shyness, alienation, and academic performance decrement. *CyberPsychology & Behavior*, 12(6):675–679.

HuggingFace-Hub. 2024. Models on huggingface hub. https://huggingface.co/csocsci.

Jina Huh-Yoo, Afsaneh Razi, Diep N. Nguyen, Sampada Regmi, and Pamela J. Wisniewski. 2023a. "Help Me:" Examining Youth's Private Pleas for Support and the Responses Received from Peers via Instagram Direct Messages. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Hamburg Germany. ACM.

Jina Huh-Yoo, Afsaneh Razi, Diep N Nguyen, Sampada Regmi, and Pamela J Wisniewski. 2023b. "help me:" examining youth's private pleas for support and the responses received from peers via instagram direct messages. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. *arXiv preprint arXiv:2302.07636*.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1195–1204.

Joseph B. Walther and Boyd Shawn. 2002. Attraction to computer-mediated social support. In *Communication technology and society: Audience adoption and uses*, pages 153–188. Hampton Press, Cresskill, NJ.

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

Matěj Kocián, Jakub Náplava, Daniel Štancl, and Vladimír Kadlec. 2022. Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12369–12377.

Christine Kormos and Robert Gifford. 2014. The validity of self-report measures of proenvironmental behavior: A meta-analytic review. *Journal of Environmental Psychology*, 40:359–371.

Maria Koutamanis, Helen G.M. Vossen, Jochen Peter, and Patti M. Valkenburg. 2013. Practice makes perfect: The longitudinal effect of adolescents' instant messaging on their ability to initiate offline friendships. *Computers in Human Behavior*, 29(6):2265–2272.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Jaroslaw Kutylowski. DeepL Translator.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

Seth C Lewis, Rodrigo Zamith, and Alfred Hermida. 2013. Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of broadcasting & electronic media*, 57(1):34–52.

Nan Lin, Walter M. Ensel, Ronald S. Simeone, and Wen Kuo. 1979. Social Support, Stressful Life Events, and Illness: A Model and an Empirical Test. *Journal of Health and Social Behavior*, 20(2):108.

Dong Liu, Kevin B. Wright, and Baijing Hu. 2018. A meta-analysis of Social Network Site use and social support. *Computers & Education*, 127:201–213.

Weixu Lu and Keith N Hampton. 2017. Beyond the power of networks: Differentiating network structure from social media affordances for perceived social support. *New media & society*, 19(6):861–879.

Sanjana Mendu, Anna Baglione, Sonia Baee, Congyu Wu, Brandon Ng, Adi Shaked, Gerald Clore, Mehdi Boukhechba, and Laura Barnes. 2020. A framework for understanding the relationship between social media discourse and mental health. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–23.

MS Neethu and R Rajasree. 2013. Sentiment analysis in twitter using machine learning techniques. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pages 1–5. IEEE.

Alicia L Nobles, Jeffrey J Glenn, Kamran Kowsari, Bethany A Teachman, and Laura E Barnes. 2018. Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–11.

Hyun Jung Oh, Elif Ozkaya, and Robert LaRose. 2014. How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. *Computers in Human Behavior*, 30:69–78.

Douglas A. Parry, Brittany I. Davidson, Craig J. R. Sewall, Jacob T. Fisher, Hannah Mieczkowski, and Daniel S. Quintana. 2021. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, 5(11):1535–1547.

PennState. 2024. Penn state, department of statistics: Introduction to mathematical statistics.

Jaromír Plhák, Michaela Lebedíková, Michał Tkaczyk, and David Šmahel. 2023a. Web-based annotation tool for instant messaging conversations. *RASLAN 2023 Recent Advances in Slavonic Natural Language Processing*, page 3.

Jaromir Plhák, Ondřej Sotolář, Michaela Lebedikova, and David Smahel. 2023b. Classification of adolescents' risky behavior in instant messaging conversations. In *International Conference on Artificial Intelligence and Statistics*, pages 2390–2404. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pretraining. *OpenAI Technical Report*.

Rohizah Abd Rahman, Khairuddin Omar, Shahrul Azman Mohd Noah, Mohd Shahrul Nizam Mohd Danuri, and Mohammed Ali Al-Garadi. 2020. Application of machine learning methods in mental health detection: A systematic review. *IEEE Access*, 8:183952–183964.

Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29.

Gretchen M. Reevy and Christina Maslach. 2001. Use of Social Support: Gender and Personality Differences. *Sex Roles*, 44(7/8):437–459.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Marko Robnik-Šikonja and Marko Bohanec. 2018. Perturbation-based explanations of prediction models. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 159–175.

John Rooksby, Alistair Morrison, and Dave Murray-Rust. 2019. Student perspectives on digital phenotyping: The acceptability of using smartphone data to assess mental health. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Hugo Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simão, and I. Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Lloyd S Shapley et al. 1953. A value for n-person games.

Adrian B. R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(9):1426–1448.

Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Trevino, and Juan Gordon. 2013. Empirical study of machine learning based approach for opinion mining in tweets. In *Advances in Artificial Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27–November 4, 2012. Revised Selected Papers, Part I 11*, pages 1–14. Springer.

Dorothy Skierkowski and Rebecca M Wood. 2012. To text or not to text? the importance of text messaging among college-aged youth. *Computers in Human Behavior*, 28(2):744–756.

Ondřej Sotolář, Jaromír Plhák, and David Šmahel. 2024. Leveraging conceptual similarities to enhance modeling of factors affecting adolescents' well-being. In *International Conference on Text, Speech, and Dialogue*, pages 263–274. Springer.

Ondřej Sotolář, Jaromír Plhák, and David Šmahel. 2021. Towards personal data anonymization for social messaging. In *International Conference on Text, Speech, and Dialogue*, pages 281–292. Springer.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, pages 197–209. Springer.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Trans. Comput.-Hum. Interact.*, 27(5).

Lisa Collins Tidwell and Joseph B. Walther. 2002. Computer-Mediated Communication Effects on Disclosure, Impressions, and Interpersonal Evaluations: Getting to Know One Another a Bit at a Time. *Human Communication Research*, 28(3):317–348.

Sigal Tifferet. 2020. Gender Differences in Social Support on Social Network Sites: A Meta-Analysis. *Cyberpsychology, Behavior, and Social Networking*, 23(4):199–209.

Emiliano Treré. 2015. Reclaiming, proclaiming, and maintaining collective identity in the# yosoy132 movement in mexico: An examination of digital frontstage and backstage activism through social media and instant messaging platforms. *Information, Communication & Society*, 18(8):901–915.

Marion K Underwood, Samuel E Ehrenreich, David More, Jerome S Solis, and Dawn Y Brinkley. 2015. The blackberry project: The hidden world of adolescents' text messaging and relations with internalizing symptoms. *Journal of Research on Adolescence*, 25(1):101–117.

Marion K Underwood, Lisa H Rosen, David More, Samuel E Ehrenreich, and Joanna K Gentsch. 2012. The blackberry project: capturing the content of adolescents' text messaging. *Developmental psychology*, 48(2):295.

Sonja Utz and Johannes Breuer. 2017. The Relationship Between Use of Social Network Sites, Online Social Support, and Well-Being: Results From a Six-Wave Longitudinal Study. *Journal of Media Psychology*, 29(3):115–125.

Patti Valkenburg and Jochen Peter. 2011. Online communication among adolescents: An integrated model of its attraction, opportunities, and risks. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, 48:121–7.

Patti M. Valkenburg, Ine Beyens, Adrian Meier, and Mariek M.P. Vanden Abeele. 2022. Advancing our understanding of the associations between social media use and well-being. *Current Opinion in Psychology*, 47:101357.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Matthias Vogt, Ulf Leser, and Alan Akbik. 2021. Early detection of sexual predators in chats. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4985–4999.

Ge Wang, Wei Zhang, and Runxi Zeng. 2019. WeChat use intensity and social support: The moderating effect of motivators for WeChat use. *Computers in Human Behavior*, 91:244–251.

J Austin Williamson and Michael W O'Hara. 2017. Who gets social support, who gives it, and how it's related to recipient's mood. *Personality and Social Psychology Bulletin*, 43(10):1355–1377.

Thomas A Wills and Ori Shinar. 2000. Measuring perceived and received social support.

Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating passage-level relevance and its role in document-level relevance judgment. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 605–614.

Bo Xie. 2008. Multimodal computer-mediated communication and social support among older chinese internet users. *Journal of Computer-Mediated Communication*, 13(3):728–750.

Lingling Xu, Cheng Yi, and Yunjie Xu. 2007. Emotional expression online: The impact of task, relationship and personality perception on emoticon usage in instant messenger. *PACIS 2007 proceedings*, page 79.

Zhiheng Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Yuxiang Chris Zhao, Mengyuan Zhao, and Shijie Song. 2022. Online health information seeking behaviors among older adults: systematic scoping review. *Journal of medical internet research*, 24(2):e34790.

Biru Zhou, Dara Heather, Alessia Di Cesare, and Andrew G. Ryder. 2017. Ask and you might receive: The actor-partner interdependence model approach to estimating cultural and gender variations in social support: Social support seeking and provision. *European Journal of Social Psychology*, 47(4):412–428.

David Álvarez García, José Núñez, Trinidad Garcia, and Alejandra Barreiro. 2018. Individual, family, and community predictors of cyber-aggression among adolescents. *The European Journal of Psychology Applied to Legal Context*.

A IM Dataset: Annotation and Statistics

We collected a dataset of online private conversations that the participants of our research led with their counterparts, which were exported from the Messenger and WhatsApp applications. They were exported by the participants themselves, who were instructed to select a timeframe for the export, which resulted in data from 2015-09-14 to 2020-12-14. This timeframe suggests, that recent topics, such as Covid-19, are not over-represented in the data. Each of the participants (13-17 years old) conversed with multiple counterparts; the final dataset contains data authored by 2165 different users. See also the demographics in Section A.2 and the authorship distribution in Figure 8.

A.1 Segmentation

As many of the conversations were long-running, some spanning years, we divided them into smaller parts to achieve better topic separation using a threshold of 60+ minutes-long conversation pauses, which resulted in a shift of the distribution of the unit length towards shorter units. After this, the total number of conversations was 90,422 with 1,260,492 total utterances.

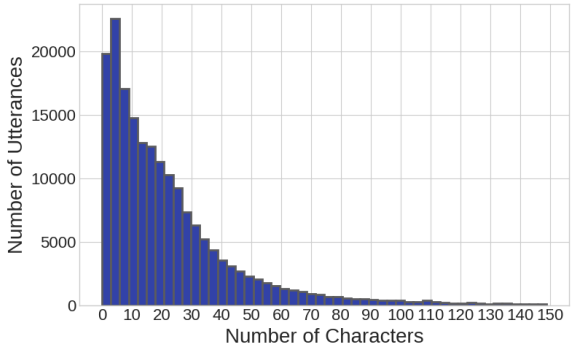


Figure 1: The distribution of utterance length in characters without outliers. 52.79% $\leq$ 20, 72.08% $\leq$ 30 characters.

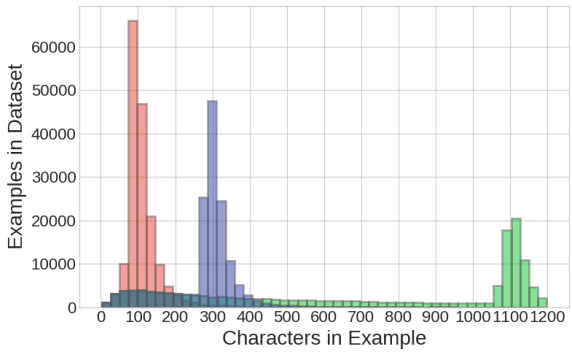


Figure 2: The distributions of example length for three settings of the context length: 64 (left), 256 (center), and 1024 (right).

A.2 Demographics

Research such as Benotsch et al. (2013); Valkenburg and Peter (2011) shows that adolescents overwhelmingly communicate with peers, which allows us to assume that most of our dataset is composed of such conversations. 87.25% of utterances are in Czech, 8,08% in Slovak (Czech and Slovak are mutually intelligible sister languages), and 4.67% in English. 53.2% of users were male and 46.79% female (see Table 2), which we determined by automatic morphological analysis using the MorphoDiTa tool from Straková et al. (2014) of the names the users chose because the anonymization method we used retains the grammatical genders. In Czech and Slovak, there is a third grammatical gender *neuter*, but we did not detect any in our data. Since we were limited by the anonymity of our data, determining other demographic factors reliably, such as nationality, location, race, religion, exact age, sexuality, and more nuanced gender, and many others, was not feasible.

Author	Utterance	Class
John soft	I'll finish the Math task tomorrow	none
John limit	Like, I really have to do it hard limit	none
Tim	The math task looks easy to me	Emotional Support
Tim	You have 6 hours to deadline, chill	Emotional Support
John	But I'm really tired after the day	none
Tim	I'm having some tea and I'm super	none

context-window labels

Figure 3: Different approaches to constructing the dataset examples. Blue: single utterance, red: hard character limit, green: soft character limit.

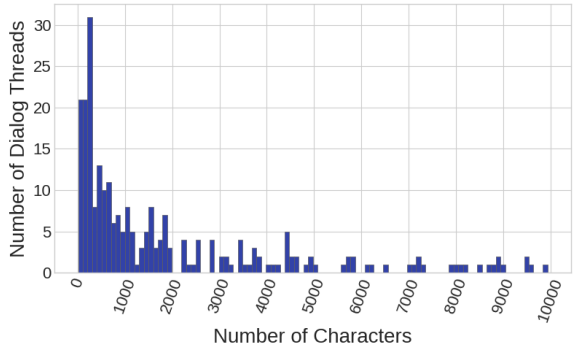


Figure 4: The distribution of dialog thread length in characters without outliers for data annotated with Supportive Interaction labels.

A.3 Annotation Process

The label definitions materialized in the annotation manual (refer to Github-Repository, 2024), which covers general and class-specific annotation rules and contains examples. We defined five labels, which are based on the five categories of supportive interactions as defined in the Social Support Behavioral Code (SSBC) developed by Cutrona and Suhr (1992) as shown in Section 1, Table 1 with the addition of the 'none' label for utterances without any instances present.

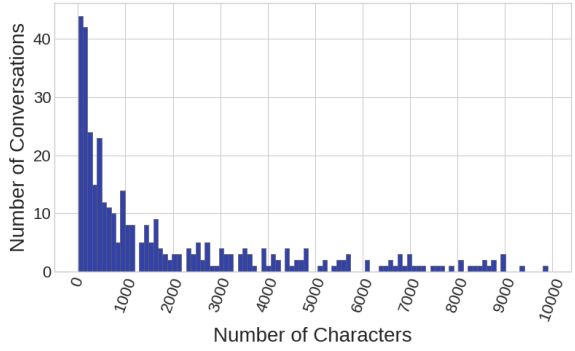


Figure 5: The distribution of conversation length (segmented using a threshold of 60+ minutes-long conversation pauses) in characters for data annotated with Supportive Interaction labels.

Each label was decided by majority vote – at least two annotators used the same label. We trained three annotators, who were research team members (male, male, and female), to apply it to samples of the data. Over several training iterations, we refined the manual until we reached the desired level of inter-annotator agreement (IAA) measured with Cohen's κ. After finalizing the manual, we threw out the initial annotations. After each example was annotated by two raters, an additional round of annotation by the supervisor to create the gold-standard version of the dataset by resolving the conflicts where labels have at least two-vote confidence. Table 3 shows the statistics for the annotations including the IAA. In Figure 6 we show the distribution of the length of annotated utterances excluding the 'none' label which follows the exponential distribution with half of the examples below 60 characters. In Figure 7 we show the distribution of the number of utterances annotated within one continuous Supportive Interaction excluding the 'none' label which is also exponential.

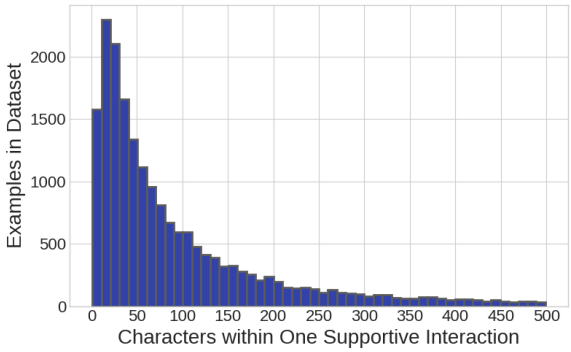


Figure 6: Distribution of the length of annotated utterances excluding the 'none' label measured in characters.

We inspect the distribution of the utterance authors in Figure 8 because it influences the reliability of the measured prediction results of our models. If the data are authored by a small number of users, the models may overfit on the vocabulary, style, and topics discussed by this small group of authors. In our data, the distribution for some of the labels is less optimal than for the others (refer to Figure 8), as more users produced higher volume of utterances than the rest, but for all the labels we consider the distribution satisfactory.

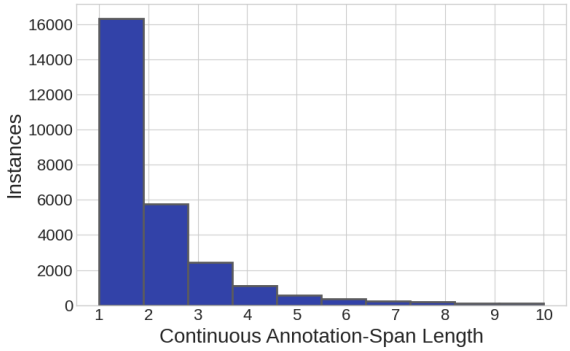


Figure 7: Distribution of the number of utterances annotated within one continuous Supportive Interaction excluding the 'none' label.

B Training Details

For the detailed training setup, see the code repository (Github-Repository, 2024).

Code https://github.com/csocsci/supportive-interactions

Models The trained models are available from (HuggingFace-Hub, 2024). Notable variants:

Fine-tuned for English, XLM-RoBERTA-large:
https://hf.co/csocsci/xlm-roberta-large-binary-en-iib

Fine-tuned for Czech, XLM-RoBERTA-large:
https://hf.co/csocsci/xlm-roberta-large-binary-cs-iib

Fine-tuned for Czech, RoBERTA:
https://hf.co/csocsci/robeczech-base-binary-cs-iib

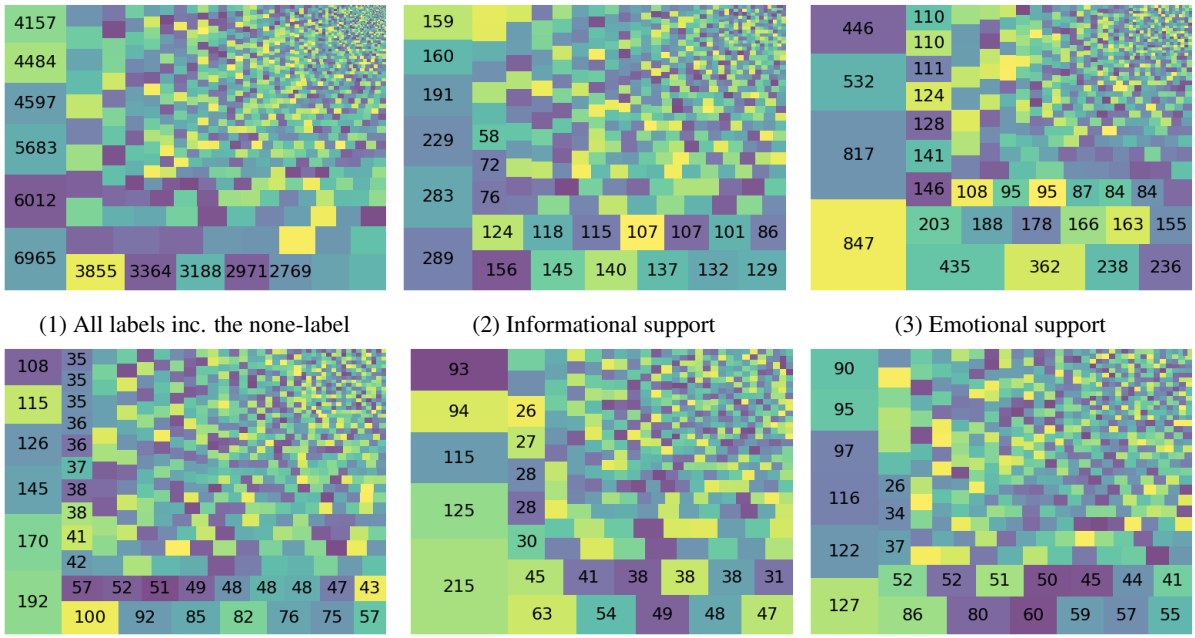
See the corresponding model cards for usage.

Training Settings We trained the models using the HuggingFace Transformers. We applied standard sequence-to-sequence training with **cross-entropy** loss on tokens. We optimized the model with **AdamW** optimizer and effective **batch size** of 256. We used **learning rate** of 5e-5 with 1000 **warmup steps**, and a **linear lr decay** to 0 in 100000 steps. The models were trained in bf16 **precision**. During training, we monitored the validation predictions on the **dev** sample and used **early stopping**. We set the convergence criteria as failing to improve the F1 over $5 * 500$ steps. After training we selected the best checkpoint. All models converged between 8000 and 15000 steps ($M_{\text{avg}} = 9500$).

Hardware To train our models, we used A100 40GB GPUs. The total training wall time, including preliminary experiments, was 17 days.

C Error Analysis

In this section, we qualitatively explore the classification results. Methods for the challenging problem of explaining predictions generated by deep neural networks can be divided into explanations by simplification, e.g., LIME (Ribeiro et al., 2016); gradient-based explanations, e.g., Integrated Gradients (Sundararajan et al., 2017); perturbation-based explanations (Shapley et al., 1953; Zeiler and Fergus, 2014). Based on reviews (Atanasova et al., 2020; Robnik-Šikonja and Bohanec, 2018), which analyzed different diagnostic properties of explainability techniques, we opted for the gradient-based technique of Layer Integrated Gradients (LIG), which outperforms the other techniques, particularly in the *Agreement with human rationales* metric, which is crucial for qualitative analysis. By using LIG, we get the contribution weight of the input tokens and highlight them in the text. We use the gradient-based LIG to get the estimated contribution weight of the input tokens resulting in Table 8.



(1) All labels inc. the none-label (2) Informational support (3) Emotional support

(4) Social companionship (5) Appraisal (6) Instrumental support

Figure 8: Utterance-author count distribution across the labels in the dataset (N = 196,772, n_2 = 7322, n_3 = 10,526, n_4 = 5766, n_5 = 2524, n_6 = 3640). One rectangle represents a user, and the relative size of it and the number within represents the number of such labeled utterances. This type of analysis shows whether the annotated phenomena are distributed across many users or just a few.

Table 8: Error Analysis of classified examples: LIG attributions with respect to input features. Note the technical limitation of the XLM-RoBERTa model family: its tokenizer incorrectly encodes some emoticons. Therefore, we can observe auxiliary characters, which replace the emoticons. The emoticons are replaced consistently; however, the replacements are void of a priori semantics from the pre-training.

Prediction	Example
TP	#s Yeah , why ? ; Will you come to the laser game with us ? ; ðŸ˜„ so yeah ; Su pr #/s
TP	#s Hel lo , are you going to the dance tonight ? ; Y ep I ' m coming ðŸ˜ ; S hall we go together ? ; Well I ' m already going with Tomas but you can join me ðŸ˜Š ; O ki â˚º ï¸ ; Me et me at my house at 12:30 . #/s
TP	#s Hey , where are you ? ; Why ? Is something wrong ? ; My stomach hurt s and I ' m sweat ing . What about you ? ; I ' m sorry , I hope you feel better ðŸ˜¬ #/s
FN	#s still im dir ty af ; Ha haha , that ' ll get you in the shower , hu h ? ; lol ; what ' s a shower ? ; https :// www . youtube . com ; No oooo o ; What no oooo ; Pro bab ly without the head phone s from ; X dd d ; It ' s not lo ud . ; É ; 10 min is enough ; So speak . ; W t f ; And what were you thinking ; Andre ych uka man has negative s ; And u need to accept it #/s
FN	#s I won ' t ; So wait a minute . ; I didn ' t laugh at your poem s ; W t f ; Th ose were my first ; The y were funny ; < PH OTO > ; Do es nt look bad lol ; O ke thank u ; I just wonder ... Why is the index finger longer than the middle finger ; 😂 ; Her e ; Lo ok at dis ; < PH OTO > ; Lo oks great ; Hey my bir d ' s gone crazy ; XD ; Yo i have to go #/s
FN	#s its my gift for him tho XD ; it was suppose to be written for him ; his idea ; and i wanted to laugh ; A haa , that ' s something else 😂 ; but then I found out that the di ary actually mirror s ur life ; if u write everything on ur mind ; what about this song ? ; https :// www . youtube . com ; I know here #/s
FP	#s Hey , what time am I supposed to come tonight ? ; So at 7 or 6 #/s
FP	#s Cau es , do you have your AJ book with you ? ; Y ep ; It ' s okay . #/s
FP	#s Cau ko , do you happen to know of any temp jobs that pay ? ; Hi , I don ' t know anything at the moment . #/s