Towards Preventing Intimate Partner Violence by Detecting Disagreements in SMS Communications

Mahesh Babu Kommalapati[1] KOMMALAPATI.M@NORTHEASTERN.EDU
Xiao Gu[4] XIAO.GU@ENG.OX.AC.UK
Harshit Pandey[1] PANDEY.HAR@NORTHEASTERN.EDU
Christie J. Rizzo[2] C.RIZZO@NORTHEASTERN.EDU
Charlene Collibee[3] CHARLENE_COLLIBEE@BROWN.EDU
Silvio Amir[1] S.AMIR@NORTHEASTERN.EDU
Aarti Sathyanarayana[1,2] A.SATHYANARAYANA@NORTHEASTERN.EDU

[1] *Khoury College of Computer Sciences, Northeastern University, Boston, USA*
[2] *Bouve College of Health Sciences, Northeastern University, Boston, USA*
[3] *Warren Alpert Medical School, Brown University, Providence, USA*
[4] *Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom*

Abstract

Intimate Partner Violence (IPV) among adolescents is a major public health concern, particularly for justice-involved adolescents which are at higher risk of experiencing IPV. Early detection of disagreements between romantic partners can provide an opportunity for just-in-time interventions to prevent escalation. Prior work has proposed methods for early detection of disagreements based on metadata features of text message conversations. In this work, we build on these prior efforts and investigate the impact of explicitly modeling the contents of text conversations for disagreement detection. We develop and evaluate supervised classifiers that combine metadata features with sentiment and semantic features of texts and compare their performance against few-shot learning with instruction-tuned Large Language Models (LLMs). We conduct experiments on a dataset collected to study the communication patterns and risk factors associated with IPV among justice-involved adolescents. In addition, we measure models' generalization to out-of-distribution samples using an external dataset comprising adolescents enrolled in child welfare services. We find that: (i) text-based features improve predictive performance but do not help models generalize to other populations; and (ii) LLMs struggle in this setting but can outperform supervised classifiers in out-of-distribution samples.

Keywords: Natural Language Processing, Large Language Models, Machine Learning, Disagreement Prediction, Juvenile Justice, Ecological Momentary Assessment

Data and Code Availability Data collection for this study has just concluded, and the data is not yet prepared for public deposit. We plan to make it available in the future. Our code is available at https://github.com/The-SATH-Lab/IPV-Disagreements

Institutional Review Board (IRB) This study involved human subjects, and we received approval from the Northeastern University Institutional Review Board(numbers are 19-03-07, 18-03-10). Informed consent was obtained from all study participants, and we performed secondary analysis on the de-identified data.

1. Introduction

Intimate Partner Violence (IPV) among adolescents is a major public health concern, particularly for youth involved with the juvenile justice system. These individuals are at a higher risk of experiencing IPV, which can lead to long-term mental health challenges, such as depression and substance abuse, and can result in repeated offenses that deepen their entanglement with the justice system (Arrojo et al., 2024; Bell and Bailey, 2021). With millions of adolescents under juvenile court jurisdiction, the need for effective early intervention strategies is paramount.

Conflicts and disagreements within romantic relationships often serve as early warning signs of IPV (Giordano et al., 2023). Early, targeted interventions have proven to be effective in reducing aggression and preventing abuse in this high-risk population (Crooks et al., 2019; DuPont-Reyes et al., 2019). However, current interventions, such as cognitive behavioral therapy, are often insufficient to prevent the escalation of conflicts in real-time. On the other hand, the widespread use of mobile technology by adolescents provides a new avenue for studying these dynamics. Text messaging, a primary mode of communication, offers a naturalistic setting for understanding relationship conflict. Prior research has demonstrated the feasibility of using SMS data and Ecological Momentary Assessment surveys (EMA) to capture real-time social behaviors among at-risk youth, including those involved in the juvenile justice system (Garcia et al., 2014; Shiffman, 2000). Such approaches can be instrumental for developing tools to support just-in-time interventions for conflict resolution.

Along these lines, Pandey et al. (2023) found that metadata features from text message communications (e.g., timing and frequency) between justice-involved adolescents and their romantic partners could be leveraged to identify disagreements. However, their proposed models did not account for the textual contents of the messages. Building on these insights, this study asks whether (and to what extent) modeling the *contents* of text messages improves the performance and generalization of disagreement detection models. To that end, we develop supervised text classifiers that combine semantic and metadata features and compare their performance against few-shot learning methods with instruction-tuned Large Language Models (LLMs).

The primary objectives of this work can be summarized as follows:

1. Assess the feasibility of predicting romantic disagreements between justice-involved adolescents using text messaging data;

2. Evaluate the impact of combining semantic and metadata features on the performance supervised classifiers;

3. Contrast the performance of supervised classifiers with that of LLMs;

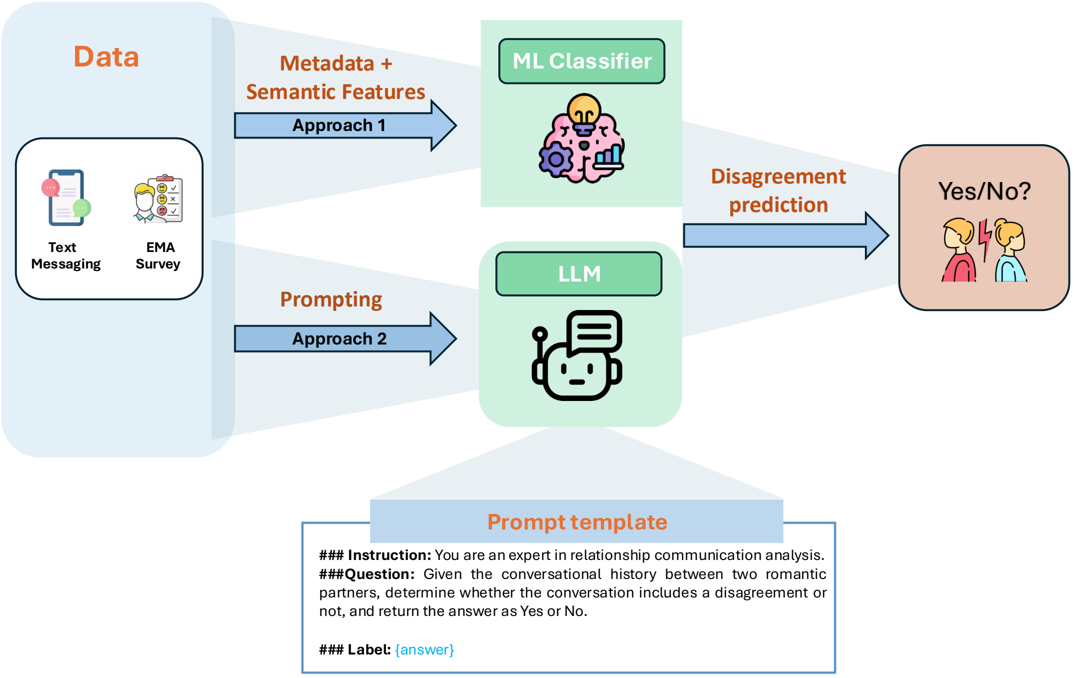
4. Validate the models' robustness to distribution shifts by assessing their generalizability to similar at-risk populations.

2. Related Work

Predicting Disagreement in Adolescents A growing body of research has demonstrated the potential of digital communication data in understanding conflict-related behaviors among adolescents. Systems such as the Youth Ecological Momentary Assessment System (YEMAS) have proven effective in capturing real-time thoughts, feelings, and behaviors through SMS-based data collection among adolescent populations in developed countries (Garcia et al., 2014). Similarly, Ecological Momentary Assessments (EMA) have been shown to be a feasible and reliable tool for studying antisocial behaviors in high-risk adolescents, including incarcerated juvenile offenders, with high rates of participation and data completion (Pihet et al., 2017). These efforts highlight the utility of mobile-based assessments in documenting behavior and conflict dynamics in at-risk youth.

Despite these advancements, prior work has primarily focused on general antisocial behaviors or abusive text detection, often relying exclusively on metadata or message content. However, limited research has explored the integration of Natural Language Processing (NLP) techniques with metadata to predict romantic disagreements, particularly among justice-involved adolescents. Our study aims to address this limitation by combining text messaging metadata and NLP techniques to develop predictive models for identifying disagreements within romantic relationships.

Ecological Momentary Assessment (EMA) Ecological Momentary Assessment (EMA) (Shiffman, 2000) has emerged as a valuable research methodology for capturing real-time data on psychological and behavioral processes in naturalistic settings. EMA leverages electronic devices, such as smartphones and wearable sensors, to collect data from participants at multiple points throughout their daily lives. By capturing data in real-time, EMA offers a more accurate understanding of the factors that influence variations in individuals' experiences, behaviors, and emotions. Our secondary analysis employs data from a study that periodically asked participants a series of questions related to their romantic relationships and emo-



Instruction: You are an expert in relationship communication analysis.
###Question: Given the conversational history between two romantic partners, determine whether the conversation includes a disagreement or not, and return the answer as Yes or No.

Label: {answer}

Figure 1: Overview of our experimental framework. We compare the performance of disagreement detection models based on supervised classifiers with metadata and semantic features against that of instruction-tuned LLMs prompted to predict disagreements given only the SMS conversations.

tional states. The survey included questions about stress levels, substance use, and various aspects of romantic relationships, such as the number and initials of current romantic partners, relationship quality, and emotions like love, satisfaction, jealousy, annoyance, obsession, and fear. Additionally, participants were asked about recent disagreements with a romantic partner.

Natural Language Processing in Mental Health Natural Language Processing (NLP) techniques has been widely applied in the field of mental health to analyze linguistic patterns in digital communication, enabling the detection of conditions such as depression, anxiety, and relational stress. Techniques such as sentiment analysis, emotion detection, and the extraction of linguistic features have proven effective in identifying mental health indicators from text data, including social media posts and text messages (Arcan et al., 2024).

3. Data

Following Pandey et al. (2023), we used the SNAP dataset to train and evaluate our disagreement prediction models. We then evaluated their robustness to out-of-distribution data by testing the performance on the SOCIAL dataset. The remainder of this section describes these datasets and Table 1 provides summary statistics.

Table 1: Descriptive Statistics of SNAP and SOCIAL Experimental Data.

	SNAP	**SOCIAL**
Participants	36	13
Average Days of Observations	13	10
Non-Disagreements instances	87	106
Disagreements instances	22	34

3.1. SNAP Dataset

The SNAP dataset was collected as part of a broader research initiative aimed at understanding the communication patterns and risk factors associated with dating violence among justice-involved adolescents (Collibee et al., 2022). The study focused on analyzing two primary types of data: text messages and Ecological Momentary Assessment (EMA) survey responses, from a population of

non-incarcerated, court-involved adolescents aged 14 to 18, who were in a romantic relationship. The population mostly reflects the demographic characteristics of youth involved with an urban family court system in the Northeastern US, however, it is not representative of the broader US population.

Our analysis utilized text messages and EMA data from **36** justice-involved adolescents over a total of 109 days, with an average of **972** messages per participant per day and an average of **13** days of data per participant. The text messages contain information such as timestamps, content, format (i.e., text, image, video, etc.), and the type of relationship with the interlocutor (i.e., Romantic Partner, Friend, Parental Figure, and Others). Non-text message formats were excluded and missing values in the EMA responses were discarded.

The surveys included questions about daily disagreements with romantic partners, providing a gold standard for identifying moments of disagreement. The following are examples of the questions asked to participants:

1. Have you had a disagreement with a romantic partner today?
2. Did you and a romantic partner have a fight/argument/disagreement today?

The days on which the participant responded positively to at least one of the questions were identified as instances of disagreement. The resulting dataset is quite imbalanced, with 78% of instances labeled as non-disagreements and only 22% as disagreements.

3.2. SOCIAL Dataset

The SOCIAL dataset involves youth in the child welfare system, recruited through families attending Family Court due to concerns such as abuse or neglect. Unlike the SNAP dataset, which focuses on youth in the juvenile justice system, SOCIAL participants represent a distinct but similarly at-risk population. Our analysis utilized text messages and EMA data from **13** participants over a total of 140 days, with an average of **967** messages per participant per day and an average of **10** days of data per participant. Similarly to the SNAP dataset, we used EMA survey responses to identify instances of disagreement and applied the same preprocessing steps. Again, we found the resulting dataset to be rather imbalanced (with 75 % of the instances showing no disagreements).

4. Modeling the Language of Romantic Disagreements

We develop text classifiers to predict disagreements between romantic partners based on the contents of SMS conversations. We consider two types of approaches: (i) Standard supervised classifiers with text features extracted from individual messages and then aggregated on a daily basis (Section 4.1); and (ii) Instruction-tuned Large Language Models prompted to determine whether a disagreement has occurred given a sequence of text messages (Section 4.2).

4.1. Supervised Text Classifiers

The supervised classifiers combine metadata features with two groups of content-based features: (i) sentiment based features; and (ii) semantic features derived from contextualized word embeddings.

4.1.1. Sentiment Features

Sentiment analysis was employed to assess the emotional tone of text messages exchanged between participants. We leveraged `TextBlob`[1], an off-the-shelf python package, to assign a polarity score $P_i \in \{-1, 0, 1\}$ to each text message i, denoting negative, neutral and positive sentiment, respectively.

These sentiment scores were used to calculate a sentiment feature encoding the proportion of messages with negative sentiment as:

$$\text{sentiment} = \frac{\sum_{i=1}^{n} \mathbb{I}(P_i = -1)}{n}$$

where n is the total number of messages sent by a participant in a day, and $\mathbb{I}$ is the indicator function, which equals 1 when the condition is met (i.e., the sentiment is negative).

[1] https://textblob.readthedocs.io/en/dev/

4.1.2. Embedding Features

To capture the semantics of the text messages, we utilized contextualized embeddings induced by DistilBERT (Sanh et al., 2020), a Pretrained Transformer Encoder available in the Huggingface library[2]. Text messages were tokenized and passed through the encoder model to extract latent representations from the last hidden state for each token in the sequence. Each token k was represented as a 768-dimensional vector $\mathbf{e}_k$ and then mean pooling was used to obtain a single vector representation for the entire message as

$$\mathbf{e}_{\text{message}} = \frac{1}{m}\sum_{k=1}^{m} \mathbf{e}_k$$

where m denotes the message length.

For participants with multiple messages in a day, a single vector summarizing their communication was created by averaging the embeddings of all messages for that day. Formally, let $\mathbf{e}^i_{\text{message}}$ be the message embedding for the i-th message of a participant on a given day. The aggregated daily embedding $\mathbf{e}_{\text{daily}}$ is computed as:

$$\mathbf{e}_{\text{daily}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{e}^i_{\text{message}}$$

where n is the total number of messages sent by a participant on that day.

4.1.3. Metadata Features

We replicated the metadata features proposed by Pandey et al. (2023) to be used as a baseline and subsequently augmented with the aforementioned text-based features. Specifically, we implemented the following metadata features:

1. **Average Word Count**: This feature quantifies the volume of communication by calculating the mean number of words exchanged between participants on a daily basis.

2. **Average Response Time**: This feature captures the temporal dynamics of the communication by measuring the average time elapsed between sent and received messages. The response time was calculated as follows:

   - Group consecutive sent messages without received responses into a *sent message stream.*

   - Group consecutive received messages without sent responses into a *received message stream.*

   - Calculate the time difference between the first sent message and the first received message in each *sent-received* stream pair.

   - Average the time differences from all sent-received pairs within a day to compute the daily average response time for each participant.

3. **Message Count**: This feature tracks the total number of messages exchanged within a given time frame, providing insights into communication frequency.

Note that unlike Pandey et al. (2023)'s approach, which operated on time-series to make predictions at the message level, our models predict disagreements between the participants given all the messages exchanged on a given day.

4.2. Large Language Models

Modern instruction-following Large Language Models (e.g, GPT-4 (OpenAI, 2024), LlaMA-3 (Llama Team, 2024) and Mistral (Jiang et al., 2023)) are able to perform a variety of tasks given only instructions, i.e., without being explicit trained to do so. We evaluated the performance of prompting LLMs to predict disagreements in both zero- and few-shot settings. This involved presenting the model with the entire day's conversational exchanges between the participant and their romantic partner. Each conversation was structured to include both "Sent" and "Received" labels, indicating the direction of each message. The model was then prompted to classify whether a disagreement occurred during that day based solely on this input.

In the zero-shot setting, we provided a detailed instruction explaining the task along with the input to be classified (see Appendix A for the full instructions). For the few-shot setting, we experimented with both **2-shot** by including one exemplar per class; and **4-shot** settings with two exemplars per class. The exemplars were sampled uniformly at random and kept fixed across all predictions.

[2] https://huggingface.co/docs/transformers/en/model_doc/distilbert

5. Experimental Evaluation

We conducted experiments to assess both the performance and generalization of our predictive models with the SNAP and SOCIAL datasets described in Section 3. The models were first trained and evaluated on the SNAP dataset which was divided into training and test sets using a temporal split. For participants with data spanning more than one day, 80% of their earliest interactions were allocated to the training set, and the remaining 20% were used for testing. Participants with only one day of data were included directly in the test set. This approach enables the models to be trained on historical data and evaluated on future interactions, providing a realistic assessment of their predictive performance. The models were subsequently evaluated on the SOCIAL dataset, which was not seen during training, to gauge their generalizability to out-of-distribution data.

5.1. Text Preprocessing

The text data was cleaned to remove noise and irrelevant information, making it suitable for sentiment analysis. This involved converting the text to lowercase, expanding contractions, and replacing slang with more formal equivalents to standardize the language. Special characters, punctuation, numbers, and emojis were either removed or converted into text to ensure uniformity. Afterward, the cleaned text was tokenized into individual words, followed by lemmatization, which reduced words to their root form. This helped ensure that different forms of the same word were treated consistently during the sentiment analysis.

5.2. Models

We evaluated the performance of our proposed features using traditional supervised predictive models (i.e., Logistic Regression, Random Forests and XGBoost) and compared these with open-source instruction-following Large Language Models (i.e., LLaMA 3.1-8B (Llama Team, 2024) and Mistral-7B (Jiang et al., 2023)). Given the imbalances in the training data, for the supervised learning models, we adjusted the class weights to give more importance to the minority class, reducing the bias towards predicting the majority class. Additionally, we selected the models' hyperparameters using grid search with a 5-fold cross-validation. See Appendix B for the full list of hyperparameters that were tuned.

5.3. Evaluation

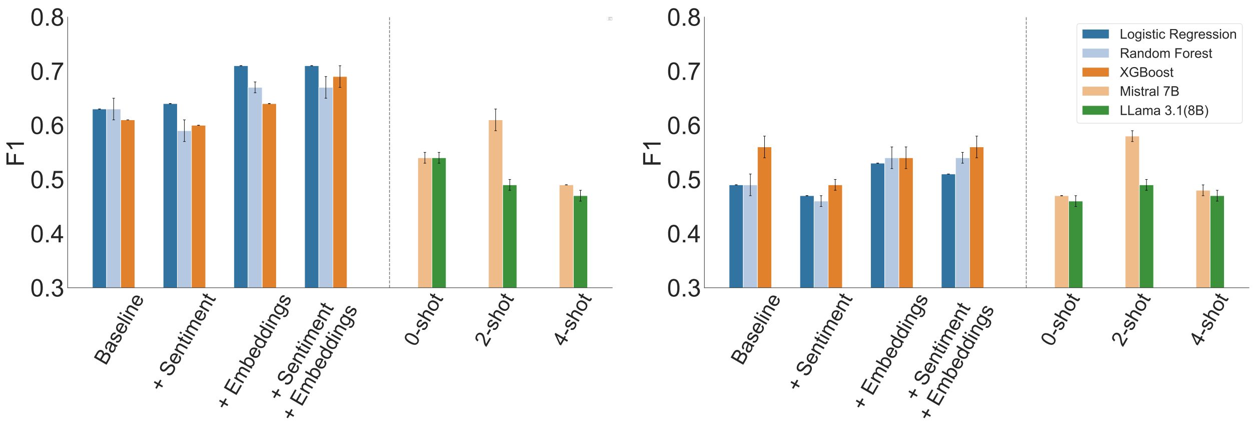
The model's performance was measured with respect to Balanced-Accuracy, Precision, Recall (for Disagreement Class), Macro F_1, and ROC-AUC. Note that computing ROC-AUC for LLMs is more challenging since these have an unconstrained output space. We addressed this issue by extracting the predicted probabilities for the tokens representing the target classes (i.e., "yes" for disagreement; and "no" for non disagreement) and renormalizing these into a probability distribution. The probability assigned to the disagreement class was then used to calculate the AUC-ROC score, allowing consistent comparison of classification performance with other models.

6. Results and Discussion

Figure 2 shows the performance of our proposed features, with respect to Macro F_1 score, for different supervised classifiers and compares these with LLMs prompted in zero and few-shot settings. The left hand side plot shows the results for the SNAP dataset (which was used to train the models) and the right hand side shows the out-of-distribution performance measured on the SOCIAL dataset. The full set of results can be found in Tables 2 and 3.

Supervised Classifiers Overall, we find that content based features improve the performance of disagreement detection classifiers, albeit to different degrees depending on the model. The sentiment feature by itself yields slight gains for Logistic Regression but it negatively impacts the other models. The semantic features, on the other hand, lead to better predictive performance across all the models (e.g., we see gains of 8% absolute F_1 for Logistic Regression). Finally, combining sentiment and semantic features provides further gains for XGBoost (of 5% absolute F_1) but does not improve over Logistic Regression and Random Forests with just the semantic features.

Regarding the out-of-distribution performance, the results are somewhat mixed (Figure 2; right hand plot). The first thing to note is that the overall performance on the SOCIAL dataset is much lower than on the SNAP data showing that models struggle to generalize to out-of-distribution samples. Second, we see that XGBoost substantially outperforms all other models with only the baseline features. The semantic features in isolation reduce model performance whereas the combination performs as well as just the baseline. For the other models, we do see



(*a*) **SNAP Test Data**

(*b*) **SOCIAL Data**

Figure 2: Comparison of F_1 score (macro) for supervised models with our proposed features and in-context learning with LLMs. Left: results on the SNAP dataset; Right: SOCIAL dataset. The Baseline approach refers to using only metadata features such as average word count and response time. Error bars represent the standard deviation of the F_1 score across different seeds for the supervised models and across different generation configurations for the in-context learning methods.

Table 2: Performance of different models and features in predicting disagreements on the SNAP test set. The Baseline refers to metadata features such as average word count and response time. For each column, the best result for each metric is **bolded**, and the second best is underlined.

	Model	Balanced Accuracy (↑)	Macro F_1 (↑)	Recall (Disagreement) (↑)	AUC Score (↑)
Baseline + Sentiment + Embeddings	Logistic Regression	**0.75** ± 0.00	**0.71** ± 0.00	0.68 ± 0.00	0.74 ± 0.00
	Random Forest	0.64 ± 0.02	0.67 ± 0.02	0.30 ± 0.04	0.74 ± 0.01
	XGBoost	0.67 ± 0.02	0.69 ± 0.02	0.41 ± 0.04	**0.76** ± 0.02
Baseline + Embeddings	Logistic Regression	0.75 ± 0.00	0.71 ± 0.00	0.68 ± 0.00	0.75 ± 0.00
	Random Forest	0.64 ± 0.01	0.67 ± 0.01	0.32 ± 0.01	0.74 ± 0.01
	XGBoost	0.62 ± 0.00	0.64 ± 0.00	0.32 ± 0.00	0.72 ± 0.00
Baseline + Sentiment	Logistic Regression	0.66 ± 0.00	0.64 ± 0.00	0.55 ± 0.00	0.74 ± 0.00
	Random Forest	0.58 ± 0.00	0.59 ± 0.02	0.27 ± 0.00	0.61 ± 0.01
	XGBoost	0.62 ± 0.00	0.60 ± 0.00	0.45 ± 0.00	0.64 ± 0.00
Baseline	Logistic Regression	0.66 ± 0.00	0.63 ± 0.00	0.55 ± 0.00	0.75 ± 0.00
	Random Forest	0.62 ± 0.02	0.63 ± 0.02	0.37 ± 0.04	0.69 ± 0.02
	XGBoost	0.64 ± 0.00	0.61 ± 0.00	0.55 ± 0.00	0.66 ± 0.00
Zero-Shot	Mistral 7B	0.59 ± 0.01	0.54 ± 0.01	0.56 ± 0.02	0.65 ± 0.00
	Llama 3.1(8B)	0.66 ± 0.02	0.54 ± 0.01	**0.80** ± 0.04	0.66 ± 0.00
2-Shot	Mistral 7B	0.61 ± 0.02	0.61 ± 0.02	0.35 ± 0.04	0.66 ± 0.00
	Llama 3.1(8B)	0.51 ± 0.01	0.49 ± 0.01	0.29 ± 0.01	0.65 ± 0.00
4-Shot	Mistral 7B	0.51 ± 0.01	0.49 ± 0.00	0.09 ± 0.02	0.64 ± 0.00
	Llama 3.1(8B)	0.50 ± 0.01	0.47 ± 0.01	0.08 ± 0.01	0.56 ± 0.00

improvements when the content features are included but these still underperform XGBoost with baseline features. These results suggest that there are significant differences in how disagreements are expressed between the participants of the SNAP and SOCIAL studies. Another possible reason for this discrepancy is that the relatively small training datasets may not be sufficient for the models to properly generalize.

Large Language Models We found that neither of the LLMs was able to perform this task in zero-shot with both models predicting a disagreement in most of the cases. However, the models performed very

Table 3: Performance of different models and features in predicting disagreements on the SOCIAL test set. This is an out-of-distribution dataset used only for testing the models. The Baseline refers to metadata features such as average word count and response time. For each column, the best result for each metric is **bolded**, and the second best is underlined.

	Model	Balanced Accuracy (↑)	Macro F_1 (↑)	Recall (Class 1) (↑)	AUC Score (↑)
Baseline + Sentiment + Embeddings	Logistic Regression	0.51 ± 0.00	0.51 ± 0.00	0.29 ± 0.00	0.50 ± 0.00
	Random Forest	0.56 ± 0.01	0.54 ± 0.01	0.13 ± 0.02	0.54 ± 0.03
	XGBoost	0.56 ± 0.01	0.56 ± 0.02	0.22 ± 0.03	0.52 ± 0.02
Baseline + Embeddings	Logistic Regression	0.53 ± 0.00	0.53 ± 0.00	0.32 ± 0.00	0.50 ± 0.00
	Random Forest	0.55 ± 0.01	0.54 ± 0.02	0.13 ± 0.02	0.54 ± 0.01
	XGBoost	0.54 ± 0.01	0.54 ± 0.02	0.18 ± 0.02	0.51 ± 0.02
Baseline + Sentiment	Logistic Regression	0.47 ± 0.00	0.47 ± 0.00	0.21 ± 0.00	0.46 ± 0.00
	Random Forest	0.49 ± 0.01	0.46 ± 0.01	0.07 ± 0.01	0.49 ± 0.02
	XGBoost	0.50 ± 0.01	0.49 ± 0.01	0.13 ± 0.02	0.49 ± 0.02
Baseline	Logistic Regression	0.49 ± 0.00	0.49 ± 0.00	0.24 ± 0.00	0.47 ± 0.00
	Random Forest	0.51 ± 0.01	0.49 ± 0.02	0.11 ± 0.02	0.48 ± 0.02
	XGBoost	0.56 ± 0.01	0.56 ± 0.02	0.22 ± 0.03	0.52 ± 0.02
Zero-Shot	Mistral 7B	0.56 ± 0.01	0.47 ± 0.00	**0.71** ± 0.02	**0.67** ± 0.00
	Llama 3.1(8B)	0.55 ± 0.01	0.46 ± 0.01	0.70 ± 0.01	0.57 ± 0.00
2-Shot	Mistral 7B	**0.57** ± 0.01	**0.58** ± 0.01	0.26 ± 0.00	0.57 ± 0.01
	Llama 3.1(8B)	0.51 ± 0.02	0.49 ± 0.01	0.13 ± 0.02	0.57 ± 0.00
4-Shot	Mistral 7B	0.50 ± 0.00	0.48 ± 0.01	0.09 ± 0.03	0.55 ± 0.00
	Llama 3.1(8B)	0.50 ± 0.00	0.47 ± 0.01	0.08 ± 0.01	0.56 ± 0.00

differently in 2-shot settings on the SNAP dataset: while Llama performed worse than zero-shot, Mistral attained an F_1 score of 0.61 (which is within the range of the baseline classifiers). Overall, we see that Mistral outperforms Llama in most of the experiments and can be competitive with the supervised classifiers (in 2-shot settings). Finally, the few-shot experiments with 4 exemplars were also unsuccessful, with models always predicting no disagreement. This may be due to the fact that the inputs for 4-shot learning were very long making them more challenging to process and interpret.

We also assessed the LLMs out-of-distribution performance by evaluating their predictions on the SOCIAL dataset using the same prompts and exemplars from the SNAP dataset. Here, we observed similar trends as in the first experiments: performance was mediocre in zero and 4-shot settings for both models, whereas in 2-shot settings Mistral attains much better performance — even outperforming the supervised baselines by 2-11% absolute F_1.

Despite the middling performance, we believe that these results are promising. Predicting disagreements from SMS conversations is a challenging task for LLMs due to the need to process very long inputs and given that is unlikely (though possible) that models have seen long SMS conversations during pre-training. As models become better at processing very long contexts it is possible that they will perform better at these tasks. This is encouraging given the difficulty of collecting large and high quality labeled datasets from vulnerable or at-risk populations.

6.1. Limitations

Despite the promising findings, several limitations must be acknowledged:

Incomplete Capture of Disagreements Disagreements between romantic partners may occur through various other forms of communication, such as in-person conversations or phone calls, which may not be captured in the text messaging data. While text messaging is a prevalent mode of communication among adolescents, it does not encompass all interactions. Our analysis focuses solely on text messages, which may lead to an incomplete understanding of the totality of disagreements occurring within relationships.

Data Representativeness The study sample consists of justice-involved adolescents from Rhode Island, which may not be representative of the broader adolescent population. The specific socio-demographic characteristics of the participants might limit the generalizability of our findings to other groups.

Reliability of Self-Reported Gold Standard This study uses a self-reported gold standard for identifying instances of disagreement, relying on participants answers to specific questions about disagreements with their romantic partners. However, self-reported data can sometimes be susceptible to bias, especially among adolescents who may feel pressure to misrepresent their responses due to influences such as fear, shame, or even the impact of a coercive partner. Future studies should consider incorporating expert assessment or alternative validation methods to strengthen the reliability of the gold standard.

7. Conclusion

In this paper, we developed and evaluated models for early detection of disagreements between justice-involved adolescents and their romantic partners from daily text message exchanges. Building on prior work that proposed using metadata features such as the timing and frequency of the messages, we assessed the impact of explicitly modeling the language of the conversations. We developed supervised classifiers combining metadata features with textual features based on (negative) sentiment and latent semantics as encoded by neural embeddings extracted from a pre-trained Transformer encoder model. These features were then aggregated to represent all the messages exchanged on given day and subsequently used to detect the occurrence of a disagreement. We compared the performance of these supervised classifiers with that of instruction-tuned LLMs (i.e., Llama and Mistral) prompted to predict whether a disagreement occurred from the text data alone.

Evaluation was first conducted on a dataset collected as part of a broader research initiative aimed at understanding the communication patterns and risk factors associated with IPV among justice-involved adolescents. In addition, we measured the models' performance on an external dataset comprising adolescents enrolled in a child welfare program, to gauge the models' ability to generalize to out-of-distribution samples. Our experimental results show that text-based features yield gains in predictive performance on in-distribution data but do not improve out-of-distribution performance as compared to metadata features alone. The results with LLMs were mixed: on the one hand, these models perform much worse than supervised classifiers on in-distribution data; on the other hand, Mistral seems to be able to generalize to out-of-distribution samples better than the supervised models.

A key limitation of this study, is the reliance on self-reported data, which can suffer from various biases such as social desirability and data missing not a random. To address this limitation, we are currently working to manually annotate the data in hopes of obtaining a larger and higher quality dataset with which to train our models. In summary, this work presents a step towards mitigating the problem of IPV that affects justice-involved adolescents. We hope that our findings can inform the design of just-in-time interventions and tools to help vulnerable and at at-risk populations combat this issue.

Ethics Statement

This study aims to enrich our the understanding of the romantic relationship dynamics and the risks of aggression among justice-involved youth. This understanding is crucial for both assessing and preventing dating aggression, which can lead to improved mental health outcomes for these adolescents.

Our research conducted a secondary analysis of personal data including EMA surveys, text messages and metadata to identify disagreements in non-incarcerated juvenile justice-involved youth. This data was collected with informed consent from participants and/or their parents or guardians and was de-identified before the study. The data was gathered non-invasively through a smartphone application running in the background and used solely for research purposes.

References

Mihael Arcan, David-Paul Niland, and Fionn Delahunty. An assessment on comprehending mental health through large language models, 2024. URL https://arxiv.org/abs/2401.04592.

S. Arrojo, F. A. Santirso, M. Lila, E. Gracia, and R. Conchell. Dating violence prevention programs for at-risk adolescents: A systematic review and meta-analysis. *Aggression and Violent Behavior*, 74:1–18, 2024. doi: 10.1016/j.avb.2023.101893. URL https://doi.org/10.1016/j.avb.2023.101893.

Tiffani L. Bell and Rahn Kennedy Bailey. Teen dating violence. In Rahn Kennedy Bailey, editor, *Intimate Partner Violence: An Evidence-Based Approach*, pages 111–113. Springer International Publishing, Cham, 2021. ISBN 978-3-030-55864-2. doi: 10.1007/978-3-030-55864-2_14. URL https://doi.org/10.1007/978-3-030-55864-2_14.

Charlene Collibee, Kara Fox, Johanna Folk, Christie Rizzo, Kathleen Kemp, and Marina Tolou-Shams. Dating aggression among court-involved adolescents: Prevalence, offense type, and gender. *Journal of Interpersonal Violence*, 37(13-14):NP12695–NP12705, 2022. doi: 10.1177/0886260521997955. URL https://doi.org/10.1177/0886260521997955. PMID: 33719683.

Claire V Crooks, Peter Jaffe, Caely Dunlop, Amanda Kerry, and Deinera Exner-Cortens. Preventing Gender-Based violence among adolescents and young adults: Lessons from 25 years of program development and evaluation. *Violence Against Women*, 25(1):29–55, January 2019.

Melissa J DuPont-Reyes, Alice P Villatoro, Jo C Phelan, Kris Painter, and Bruce G Link. Adolescent views of mental illness stigma: An intersectional lens. *Am J Orthopsychiatry*, 90(2):201–211, August 2019.

Carolyn Garcia, Rachel R Hardeman, Gyu Kwon, Elizabeth Lando-King, Lei Zhang, Therese Genis, Sonya S Brady, and Elizabeth Kinder. Teenagers and texting: use of a youth ecological momentary assessment system in trajectory health research with latina adolescents. *JMIR Mhealth Uhealth*, 2(1):e3, January 2014.

Peggy C. Giordano, Mackenzie M. Grace, Wendy D. Manning, and Monica A. Longmore. Gender, relationship concerns, and intimate partner violence in young adulthood. *Journal of family violence*, 38(4):597–609, 2023. ISSN 0885-7482. doi: 10.1007/s10896-022-00399-1. URL http://dx.doi.org/10.1007/s10896-022-00399-1.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

AI@ Meta Llama Team. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

H. Pandey, C. Rizzo, C. Collibee, and A. Sathyanarayana. Modeling messaging metadata to identify digital disagreements among non-incarcerated adolescents in the juvenile justice system. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Los Alamitos, CA, USA, sep 2023. IEEE Computer Society. doi: 10.1109/ACII59096.2023.10388170. URL https://doi.ieeecomputersociety.org/10.1109/ACII59096.2023.10388170.

Sandrine Pihet, Jill De Ridder, and Maya Suter. Ecological momentary assessment (ema) goes to jail. *European Journal of Psychological Assessment*, 33(2):87–96, 2017. doi: 10.1027/1015-5759/a000275. URL https://doi.org/10.1027/1015-5759/a000275.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Saul Shiffman. Real-time self-report of momentary states in the natural environment: Computerized ecological momentary assessment. In *The science of self-report: Implications for research and practice.*, pages 277–296. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2000.

Appendix A. Prompts

Prompt template for zero-shot learning

Instruction: Analyze the conversation between romantic partners enclosed in square brackets, determine if there is a disagreement or not, and return the answer as the corresponding label:

> [text: {text}]
> label:

Prompt template for few-shot learning

Instruction: You're an expert at analyzing conversations between romantic partners.

Go through the conversations and corresponding labels provided below to get a sense of the conversations and later classify a new conversation to determine if there is a disagreement or not, and you must return the answer as the corresponding label Yes or No. Do not add anything else.

Examples:
{{few_shot_examples}}

Analyze this conversation between romantic partners enclosed in square brackets:

[text: {text}]

Now, determine if there is a disagreement or not in the above conversation, and return the answer as the corresponding label Yes or No. Do not add anything else.

label:

Figure 3: Zero-shot, Few-shot prompt template along with conversational history.

Appendix B. Hyperparameters

Decoding Methods for Disagreement Prediction

```
greedy_search = {
    “num_beams” : 1,
    “do_sample” : False,
    “max_new_tokens” : 1
}
beam_search = {
    “num_beams” : 2,
    “do_sample” : False,
    “max_new_tokens” : 1,
    “early_stopping” : True,
}
sampling_top_k = {
    “do_sample” : True,
    “num_beams”: 1,
    “max_new_tokens”: 1,
    “early_stopping”: True,
    “temperature”: 0.2,
    “top_k”: 50
}
sampling_top_p = {
    “do_sample” : True,
    “top_k”: 0,
    “num_beams”: 1,
    “max_new_tokens”: 1,
    “early_stopping”: True,
    “temperature”: 0.3,
    “top_p”: 0.9
}
general_sampling = {
    “do_sample” : True,
    “top_k”: 50,
    “num_beams”: 1,
    “max_new_tokens”: 1,
    “early_stopping”: True,
    “temperature”: 0.5,
    “top_p”: 0.9
}
```

Figure 4: Decoding methods used to obtain disagreement label on participants daily conversational data.

Hyper-parameters used for Supervised Classifiers

Logistic Regression

```
param_grid_lr = {
   "penalty": ["l1", "l2"],
   "max_iter": [100, 200, 300]
}
```

Random Forest Classifier

```
param_grid_rf = {
   "n_estimators": [50, 100, 200, 500],
   "max_depth": [None, 10, 20],
   "min_samples_split": [2, 5, 10],
   "min_samples_leaf": [1, 2, 4]
}
```

XGBoost Classifier

```
param_grid_xgb = {
   "n_estimators": [100, 200, 500],
   "max_depth": [3, 5],
   "learning_rate": [0.1, 0.2],
   "subsample": [0.8, 1.0],
   "colsample_bytree": [0.8, 1.0],
   "gamma": [0, 0.1],
   "scale_pos_weight": [1, 3, 5]
}
```

Figure 5: Experimented with these set of Hyper-parameters to obtain best configuration.