Multilingual Serviceability Model for Detecting and Ranking Help Requests on Social Media during Disasters

FEDOR VITIUGIN @vitiugin, Universitat Pompeu Fabra, Spain
HEMANT PUROHIT @hemant_pt, George Mason University, USA

Code and data
https://shorturl.at/E3Jen

INTRODUCTION

- Emergency services face difficulties in detecting and prioritizing critical requests from social media users due to the overwhelming amount of information and limited human resources during mass emergencies or disaster events.
- While recent studies have focused on characterizing and automatically detecting help requests on social media, they focused on non-behavioral features and monolingual data, primarily in English.

Event	Serviceability	Message
Turkey-Syria Earthquake 2023	serviceable (help request)	@SERVICE Hayrullah mahallesi 16. sokak'taki Ferhat apartmanında acil yardıma ihtiyaç var! **EN:** @SERVICE Urgent help needed at Ferhat apartment in 16th street, Hayrullah neighborhood!
Hurricane Sandy 2012	serviceable (information request)	@SERVICE how I can volunteer to help clean up after the hurricane?
Catalonia Fires 2019	non-serviceable (gratitude and complaining)	@SERVICE Realizáis un gran trabajo y no os pagan lo suficiente por ello, de verdad muchas gracias **EN:** @SERVICE You guys do a great job, and you don't get paid enough for it, really thank you so much

Table 1. Examples of multilingual messages with varied serviceability characteristics that were directed at emergency services' accounts on social media (*paraphrased for anonymity*).

RESEARCH QUESTIONS

1. How to train a classification and ranking model for a social media platform to process multilingual serviceable requests while learning different types of human behaviors when seeking help?
2. To what extent does the performance improvement of the proposed framework depend on the types of behavior-guided models used?
3. Are there any differences in attention on various parts of a request content resulting from behavioral fine-tuning, to analyze the model's understanding of relevant human behaviors in multilingual requests?
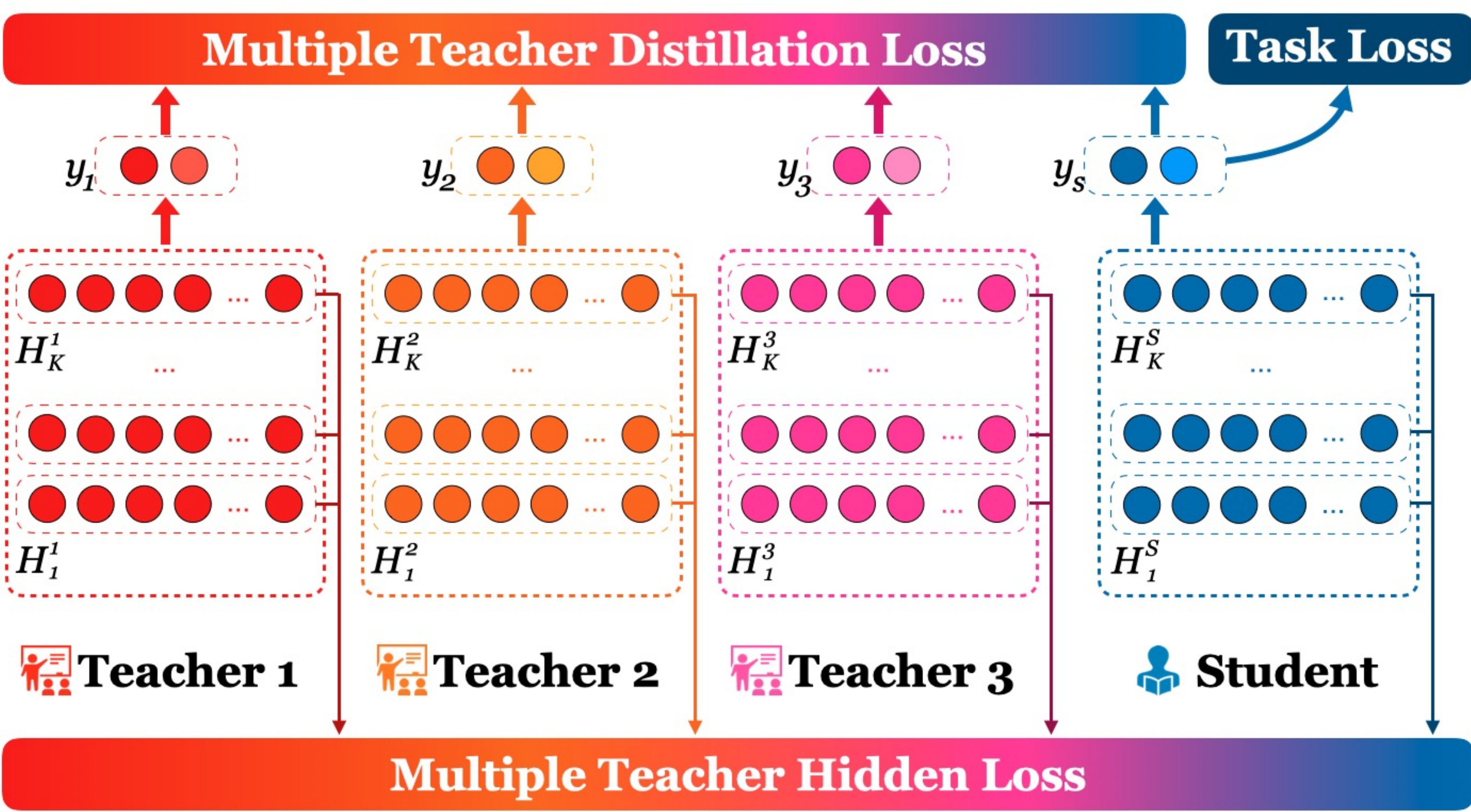
APPROACH

Figure 1. The overall architecture of MulTMR.

DATA COLLECTION

We considered ten disaster events between October 2012 and February 2023. Dataset contains messages in English, Spanish, and Turkish languages posted during disasters in seven countries.

Behavior type	Example	Type
Imperative mood requests	Prohibit rockets, firecrackers, and dangerous activities with…	False Negative
Imperative mood requests	Many of us have no choice... I have to drive 160km to go to work	False Positive
Sarcastic questions	Now if you want the Spanish army to come in and get your chestnuts out of the fire, right?	False Positive
Sarcastic questions	What does it say? Sorry but I don't speak Catalan and I want to find out	False Negative
Short question	Is there no more fire?	False Negative
Information requests in a form close to complain	You say we are strong together, you prevent aid. You cannot rule alone. People die while waiting for instructions. There are voices coming from under the buildings but you are passing by.. Is this unity????	False Negative
Contextual requests	Gazi Mustafa Kemal street No:50/A Opposite Güneşli mosque Elbistan, Kahramanmaraş	False Negative

Table 2: Erroneous examples in preliminary analysis (*paraphrased for anonymity*).

FINDINGS

- The utilization of behavior-guided models improve a Multiple Teacher-Student model in reducing uncertainty of results produced by a task-related teacher model alone.

| Model Scheme	ACC	F1	AUC
LSTM+DistilmBERT	80.92 ± 0.65	62.21 ± 5.74	85.74 ± 3.40
BERT	80.85 ± 2.38	81.19 ± 2.09	80.04 ± 0.99
XLM-RoBERTa	82.69 ± 1.35	83.04 ± 1.09	82.69 ± 1.39
* MulTMR-BERT	88.59 ± 1.87	88.76 ± 1.76	88.04 ± 1.50
* MulTMR-RoBERTa	**88.97 ± 1.23**	**89.07 ± 1.19**	**88.23 ± 1.16**

Table 3. 5-fold Cross Validation (CV) results of the binary classification task. The best performances are in bold. Models were trained on multilingual data: train (67%) – validation (13%) – test (20%). * — denotes the proposed models.

- Extensive analyses shows the value of modeling with multiple teachers, which can help adapt the model to different languages, event types, and tasks.

	baseline	+ questions	+ sarcasm	all-teachers
ACC	80.85	87.82	87.94	88.59
Compare:				
- baseline (target)	-	0.00021	**0.00017**	0.00014
- all-teachers	-	0.4244	0.5225	-
F1	81.19	87.83	88.03	88.76
Compare:				
- baseline (target)	-	0.00013	**0.00010**	0.00008
- all-teachers	-	0.3191	0.4099	-
AUC	80.04	85.91	86.90	88.04
Compare:				
- baseline (target)	-	0.00001	**0.00000**	0.00000
- all-teachers	-	0.02878	0.4268	-

Table 4. The impact of Behavior-guided models in MulTMR model is studied for the model schemes with the different teachers. The table shows p-value for each scheme's performance in comparison with the baseline (BERT) and all-teachers model (MulTMR-BERT).

- MulTMR pays close attention to important parts in the context and learns to give higher attention to the key elements of behavioral characteristics in serviceable requests classification and ranking.
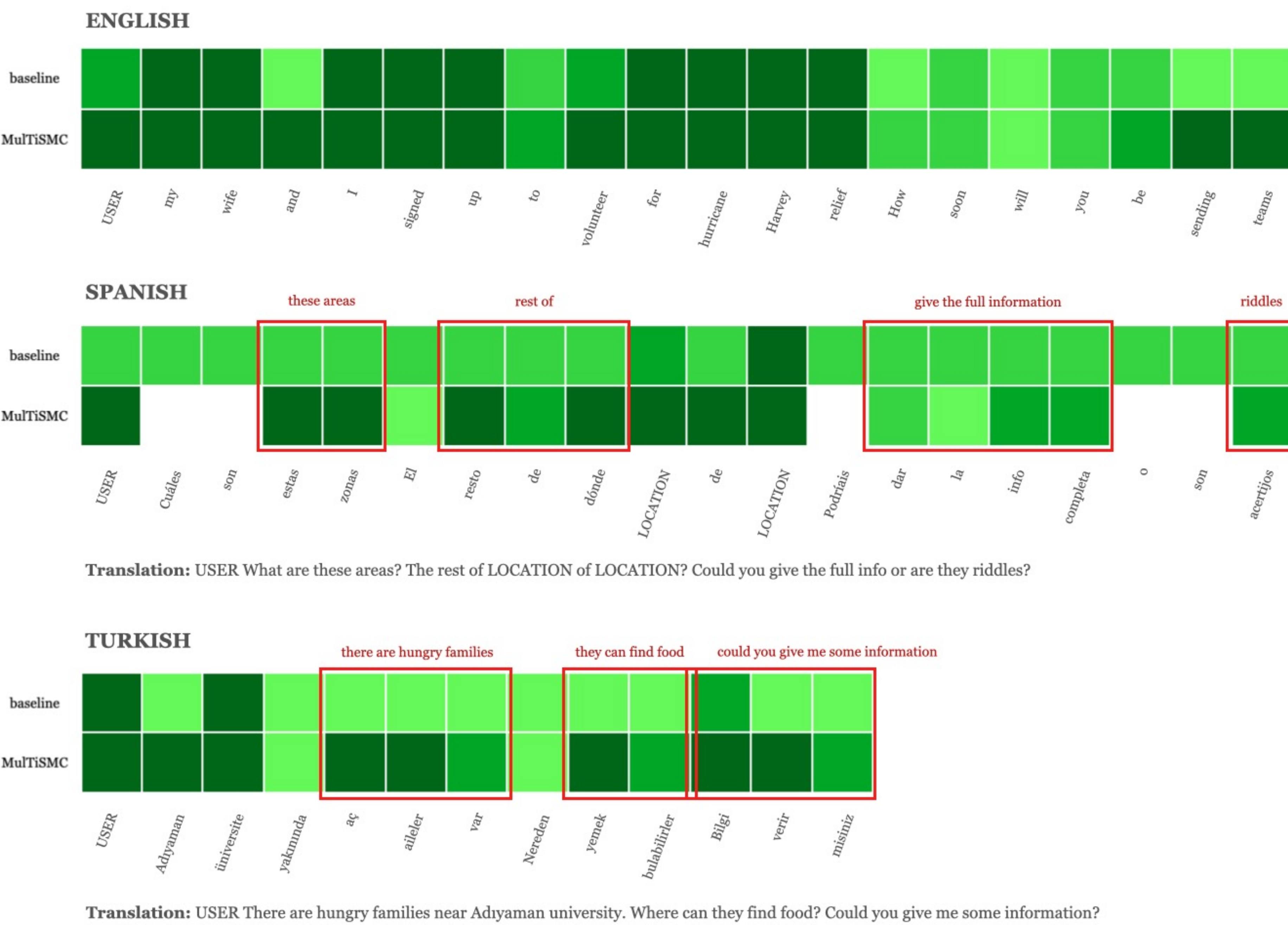
Figure 2. Attention weight maps of the texts in English, Spanish, and Turkish. The darker color indicates the higher weight (*paraphrased for anonymity*).

LIMITATIONS

- Set of languages in datasets and language models
- Relevance of experiment data

REFERENCES

[1] Kim, Y. J.; and Hassan, H. 2020. FastFormers: Highly Efficient Transformer Models for Natural Language Understanding. In Proc. of SustaiNLP.
[2] Wu, C.; Wu, F.; and Huang, Y. 2021. One Teacher is Enough? Pre-trained Language Model Distillation from Multiple Teachers. In Proc. of ACL-IJCNLP.
[3] Purohit, H.; Castillo, C.; Imran, M.; and Pandey, R. 2018. Social-eoc: Serviceability model to rank social media requests for emergency operation centers. In Proc. of ASONAM, IEEE.