Help! Need Advice on Identifying Advice

EMNLP 2020

Venkata S Govindarajan[1], Benjamin T Chen[2], Rebecca Warholic[3], Katrin Erk[1], Junyi Jessy Li[1]

[1] The University of Texas at Austin
[2] Amazon Inc.
[3] McGill University

Parenting with a history of depression?

① I took my meds the whole time. I used the tools I learned in therapy. I talked on Reddit with others to get support and ideas.

(r/AskParents)

People often give advice **implicitly** using personal narratives and other strategies (Abolfathiasl et al. 2013).

Is it too late to start a hobby/activity at 12?

② ..you can always pick anything up you think is interesting and giving it a shot. You never know what you are good at until you try new things! Idk if you have a budget or maybe borrow tools but you can try woodworking? It's fun and frustrating (in a good way) at the same time

(r/needadvice)

Advice is often **interspersed** with support, reassurance and reasoning.

How do people give advice (online)?

Advice Questions Dataset of advice-seeking intentions from personal narratives (Fu et al. 2019).

Suggestion Mining SemEval-2019 introduced a pilot task on suggestion mining but suggestions are not synonymous with advice (Negi et al. 2019).

TuringAdvice A framework that evaluates language models by asking them to generate useful advice for humans (Zellers et al. 2020).

OUR QUESTIONS

How is advice structured online?

This work aims to advance both our understanding of how people give advice, as well as to provide resources for learning to identify advice

How good are computational models at identifying advice?

We establish preliminary baselines with rule-based models (Negi et al. 2019, Potamias et al. 2019) and BERT (Devlin et al. 2019), and analyze their performance.

ANNOTATION PROTOCOL AND DATASET

DATA SOURCES

To model general online human advice-seeking interactions, we chose to construct datasets from Reddit forums (subreddits) focused on advice.

r/AskParents	**r/needadvice**
parents seeking advice	a general advice forum
less moderation	more moderation
no flairs	5 flairs – “Education”, “Career”, “Mental Health”, “Life Decisions”, “Friendships”

ANNOTATION

Reply 1

In Kentucky it 's legal to leave a " mature " 8 year old at home alone all day .
I find that crazy young .

I started leaving mine home at age 9 - 10 for a half hour here , 45 min there , working up to a couple of hours .

>>>> Reply 1.1

Yeah , even though my son has always been very mature for his age , I would not have been comfortable leaving him home alone all day long at age eight !

Reply 2

For an hour ?
I 'd on average say elementary school aged .
So 6 up , depending on how responsible / mature the child is and if they 're willing to stay home alone .

No answering the door , no leaving the house , no using the stove , no friends over and I 'd talk about what neighbors might be home in case of an emergency .

Oh , and I 'm from Germany .

ANNOTATION

Reply 1

Advice
In Kentucky it 's legal to leave a " mature " 8 year old at home alone all day .
I find that crazy young .

Advice
I started leaving mine home at age 9 - 10 for a half hour here , 45 min there , working up to a couple of hours .

>>>> Reply 1.1

Advice
Yeah , even though my son has always been very mature for his age , I would not have been comfortable leaving him home alone all day long at age eight !

Reply 2

Advice
For an hour ? I 'd on average say elementary school aged . So 6 up , depending on how responsible / mature the child is and if they 're willing to stay home alone .

No answering the door , no leaving the house , no using the stove , no friends over and I 'd talk about what neighbors might be home in case of an emergency .

Oh , and I 'm from Germany .

5 annotators on Amazon Mechanical Turk annotated each HIT of 5 comments.

We chose **sentences** as the units of advice.

How to aggregate sentence labels while accounting for inter-annotator variability?

Dawid-Skene Labels

An EM based algorithm that estimates the label with the maximum estimated **posterior probability** by iteratively computing annotator competencies and type probabilities (Dawid et al. 1979).

DATASET

r/AskParents	10,594 sentences	407 posts
r/needadvice	7,862 sentences	277 posts

Data (and code) available at

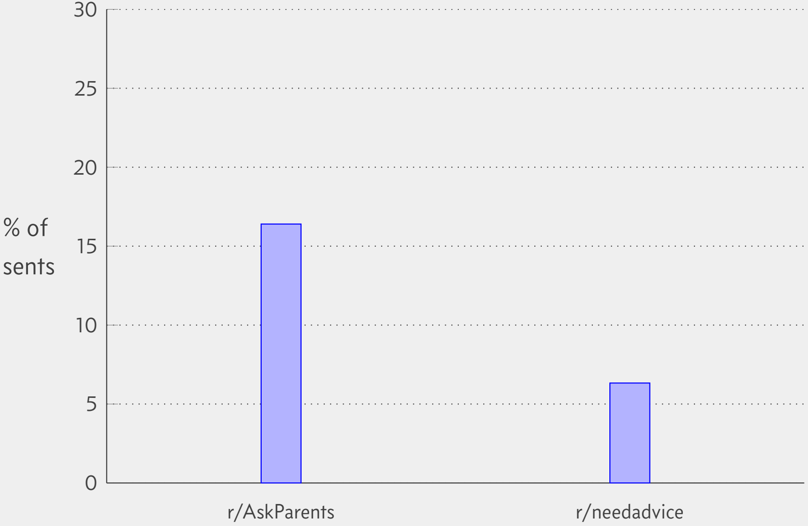
GITHUB.COM/VENKATASG/ADVICE-EMNLP2020

ADVICE STRATEGIES

③
a. I did the classic Ferberizing: check on baby after 5 mins,then 10 mins,then 20 mins, etc, until asleep. PERSONAL NARRATIVE
b. Have you tried a calm spray? QUESTIONS
c. Figure out why they like them , and then recommend those ones for those reasons. IMPERATIVES
d. If he doesn’t want therapy, maybe an antidepressant would help. CONDITIONALS

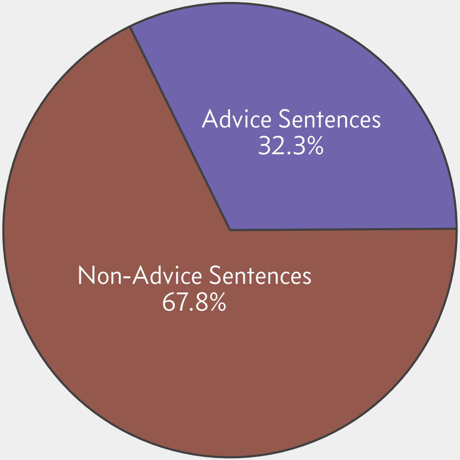
To study personal narratives further, we (the authors) analyzed 213 sentences DS-labelled as advice for whether they contained personal narratives.

PERSONAL NARRATIVES



Y Axis shows % of sentences that were judged to contain personal narratives.

NON-ADVICE



Proportion of advice and non-advice in our dataset

4

a. …being fully prepared for an interview calmed me down … **Good luck on your interviews and fingers crossed.** SENTIMENT

b. Look for smaller outfits, they're more likely to be willing to give you some time. **Most professionals - if they have the time - are more than happy to talk to a student about what they do…** SUPPORT

c. **Yes, no one will ever know the big answers to the big questions.** What is the only thing that if shared , will grow larger in size? Answer: Love. Let that define your actions in life. REASONING

	Advice	Non-advice
r/AskParents	book if take something help then you might talk need down can etc play find show or great also give buy big watch diaper car about else minute spend baby	**luck** **sorry** **shit** however dog crazy teenager op die eventually three wish weird daughter yeah brother example miss gender anyway anymore comment morning lol boyfriend girl younger **hope** drive mine
r/needadvice	he phone night adult stay set big game doctor fun bring less show love depend activity eat normal put teacher family etc minute teach allow home they area	**luck** degree company college interview hobby student field mental course op **sorry** job dog anxiety hire eventually position path **shit** comment human online community shoe **thanks** note exercise depression slowly

Top 30 lemmas ranked by log-odds ratio

MODELING

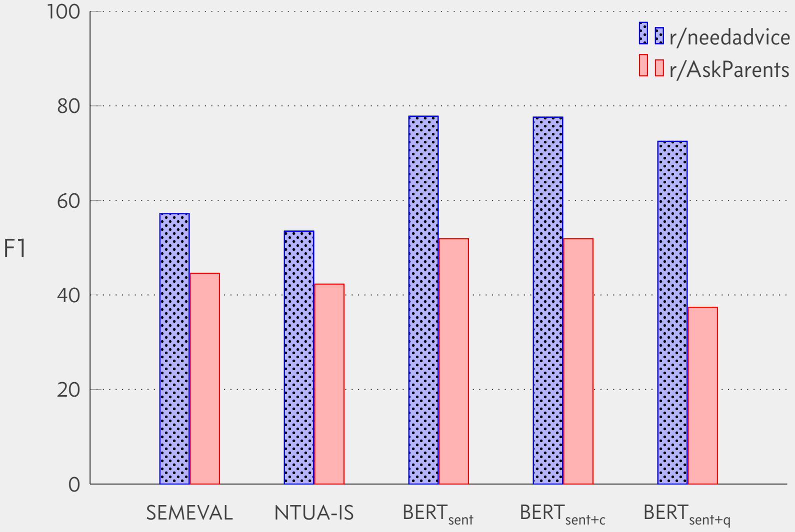
We model advice identification as a **binary classification task**.

Rule-based SEMEVAL 2019 baseline & NTUA-IS 2019 (Potamias et al. 2019).
Match and score against words, phrases, regexs:
suggest, *recommend*, *.*would\slike.*if.**

Language Models We use BERT (Devlin et al. 2019) and experiment with input:

$BERT_{sent}$	[CLS] SENTENCE [SEP]
$BERT_{sent+q}$	[CLS] SENTENCE [SEP] QUESTION
$BERT_{sent+c}$	[CLS] SENTENCE [SEP] CONTEXT

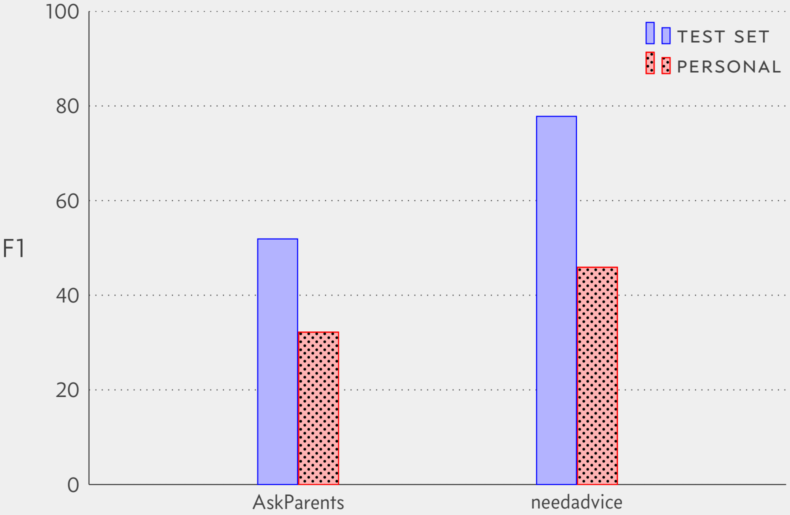
RESULTS



BERT$_{sent}$ has best performance.

Performance on r/AskParents worse than r/needadvice

PERFORMANCE ON PERSONAL NARRATIVES



$\text{BERT}_{\text{sent}}$ performance on personal narrative sentences in test set suffers.

CONCLUSION

Dataset We introduce a new dataset for advice given online.

Advice Structure People use various strategies when giving adviceaa.

Modeling Language models learn some surface-level rules, but need to do better at implicit advice.

REFERENCES

Abolfathiasl, Hossein & Ain Nadzimah Abdullah. 2013. Pragmatic Strategies and Linguistic Structures in Making 'Suggestions': Towards Comprehensive Taxonomies. *International Journal of Applied Linguistics and English Literature* 2(6). 236–241.

Dawid, A. P. & A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1). 20–28.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fu, Liye, Jonathan P. Chang & Cristian Danescu-Niculescu-Mizil. 2019. Asking the Right Question: Inferring Advice-Seeking Intentions from Personal Narratives. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 528–541. Minneapolis, Minnesota: Association for Computational Linguistics.

Negi, Sapna, Tobias Daudert & Paul Buitelaar. 2019. SemEval-2019 Task 9: Suggestion Mining from Online Reviews and Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 877–887. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Nye, Benjamin & Ani Nenkova. 2015. Identification and Characterization of Newsworthy Verbs in World News. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1440–1445. Denver, Colorado: Association for Computational Linguistics.

REFERENCES II

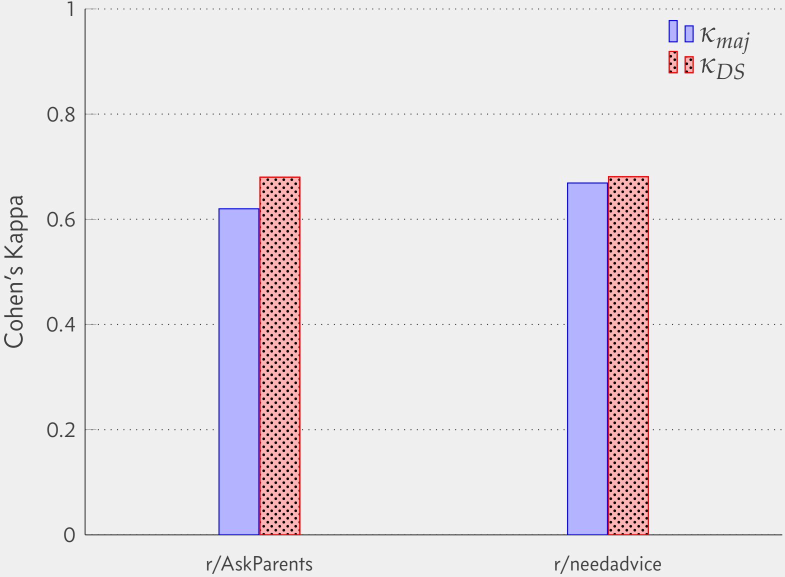
Potamias, Rolandos Alexandros, Alexandros Neofytou & Georgios Siolas. 2019. NTUA-ISLab at SemEval-2019 Task 9: Mining Suggestions in the wild. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 1224–1230. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

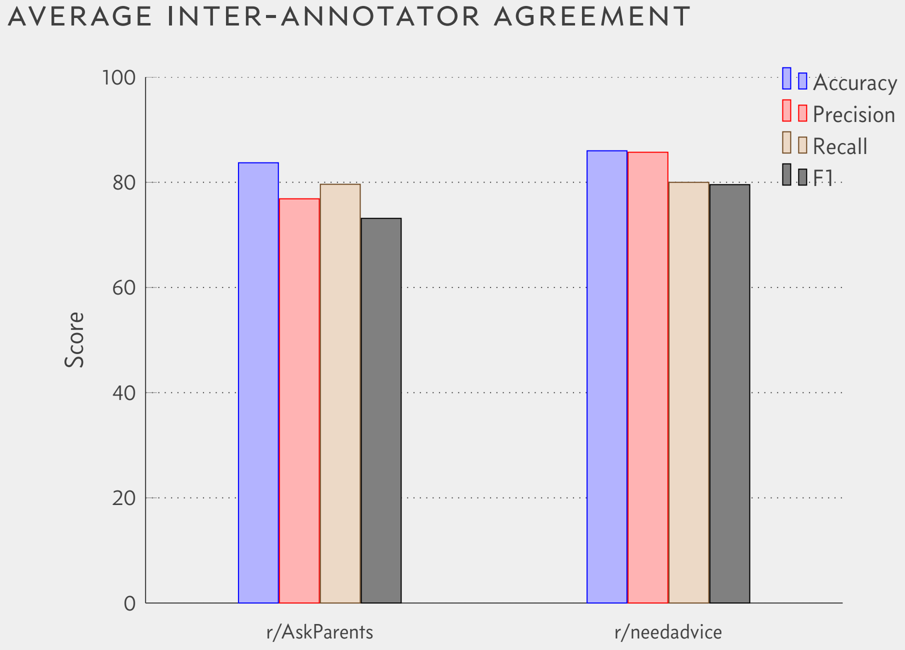
Vig, Jesse. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 37–42. Florence, Italy: Association for Computational Linguistics.

Zellers, Rowan, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi & Yejin Choi. 2020. Evaluating Machines by their Real-World Language Use. *Computing Research Repository* arXiv:2004.03607.

APPENDIX

GOLD ANNOTATOR AGREEMENT




AVERAGE INTER-ANNOTATOR AGREEMENT
Accuracy
Precision
Recall
F1
Score
100
80
60
40
20
0
r/AskParents
r/needadvice

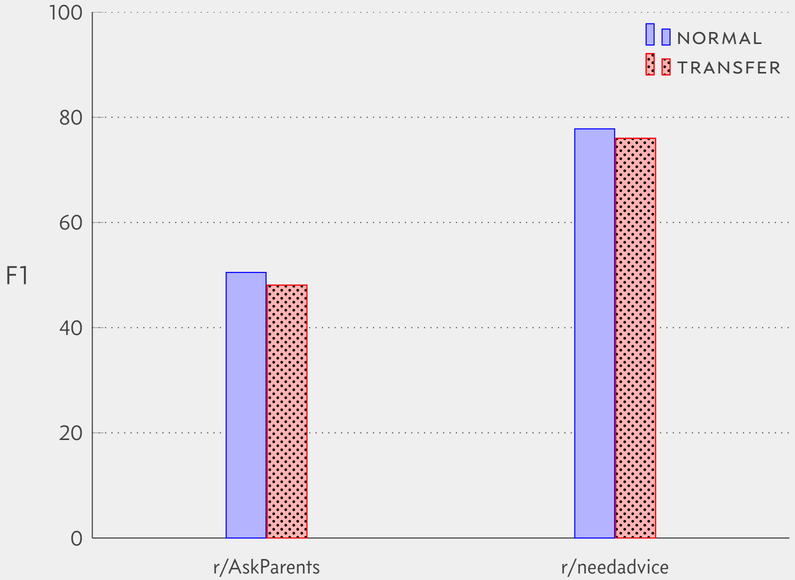
LEXICAL ANALYSIS

We quantify how strongly individual lemmas are associated with advice versus non-advice text using the log-odds ratio (Nye et al. 2015).

$$Odds(w, c) = \frac{P(w|c)}{1 - P(w|c)} \tag{1}$$

$$\text{log-odds ratio} = \frac{Odds(w, advice)}{Odds(w, non - advice))} \tag{2}$$

GENERALIZABILITY RESULTS



DATASET METRICS

Dataset	Train	Dev	Test
AskParents	8701(.29)	802(.33)	1091(.26)
needadvice	6148(.37)	816(.34)	898(.37)

Sentence metrics in our dataset, with fraction DS-labeled as advice.

GOLD INTERNAL AGREEMENT

Dataset	**Sentences**	κ_{maj}	κ_{DS}
AskParents	203	0.620	0.669
needadvice	110	0.680	0.681

Gold annotator agreement on the internal task.

AGREEMENT

Dataset	Acc	P	R	F1
AskParents	83.71	76.86	79.62	73.14
needadvice	85.99	85.71	79.99	79.55

Average inter-annotator agreement for all workers against DS labels

DISCOURSE MODES

Subreddit	Other (%)	Personal Narrative (%)
r/AskParents	83.6	16.4
r/needadvice	93.67	6.33
-Career	100	0
-Mental Health	81.82	18.18
-Friendships	100	0
-Education	95.4	4.6
-Life Decisions	88.9	11.1

Modes of discourse for advice sentences in each flair/subreddit

RESULTS-CLASSIFICATION

	Model	**P**	**R**	**F1**
r/AskParents	SEMEVAL	32.7	70.2	44.6
	NTUA-IS	31.4	64.9	42.3
	$\text{BERT}_{\text{noft}}$	62.6 (1.2)	14.9 (1.0)	24.0 (1.4)
	$\text{BERT}_{\text{sent}}$	54.9 (2.4)	49.5 (4.4)	**51.9** (1.9)
	$\text{BERT}_{\text{sent+c}}$	54.2 (2.1)	49.9 (4.0)	51.9 (2.2)
	$\text{BERT}_{\text{sent+q}}$	61.0 (13.4)	33.1 (11.9)	37.4 (8.1)
r/needadvice	SEMEVAL	44.5	80.3	57.2
	NTUA-IS	43.0	70.9	53.5
	$\text{BERT}_{\text{noft}}$	82.9 (0.5)	44.6 (1.4)	58.0 (1.2)
	$\text{BERT}_{\text{sent}}$	79.7 (3.8)	76.3 (3.9)	**77.8** (0.3)
	$\text{BERT}_{\text{sent+c}}$	80.4 (4.4)	75.3 (4.4)	77.6 (0.7)
	$\text{BERT}_{\text{sent+q}}$	83.4 (4.8)	64.7 (7.4)	72.5 (3.5)

Classification results on test set.

RESULTS-GENERALIZATION

Model	P	R	F1
AP $\longrightarrow$ AP	54.9 (2.4)	49.5 (4.4)	51.9 (1.9)
AP_p $\longrightarrow$ AP	59.1 (3.5)	44.4 (4.1)	50.5 (1.8)
NA $\longrightarrow$ AP	61.9 (4.9)	39.7 (3.5)	48.1 (1.3)
NA $\longrightarrow$ NA	79.7 (3.8)	76.3 (3.9)	77.8 (0.3)
AP $\longrightarrow$ NA	74.0 (4.0)	79.3 (2.9)	76.5 (0.9)
AP_p $\longrightarrow$ NA	76.9 (3.8)	75.5 (4.7)	76.0 (1.1)

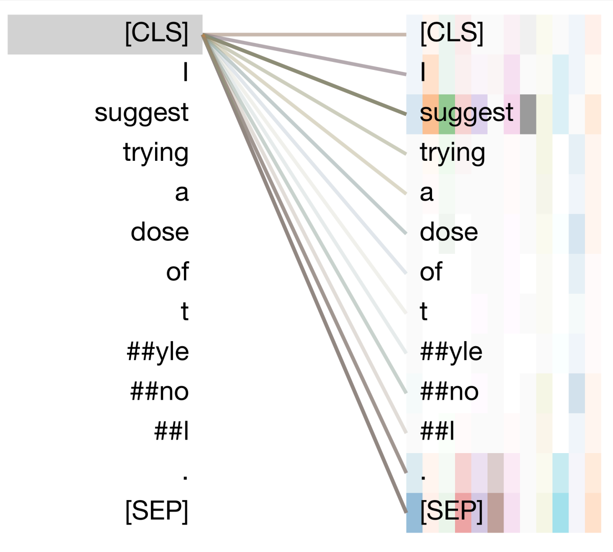
Generalizbility results on test set.

RESULTS-FLAIR

Flair	**P**	**R**	**F1**
Friendships	85.5 (5.7)	93.8 (0.0)	89.2 (2.9)
Mental Health	75.6 (3.5)	74.7 (3.6)	75.0 (0.6)
Education	86.8 (2.9)	67.4 (6.2)	75.7 (3.1)
Career	75.9 (5.1)	78.0 (3.8)	76.7 (1.3)
Life Decisions	82.4 (4.4)	82.8 (3.5)	82.4 (0.7)

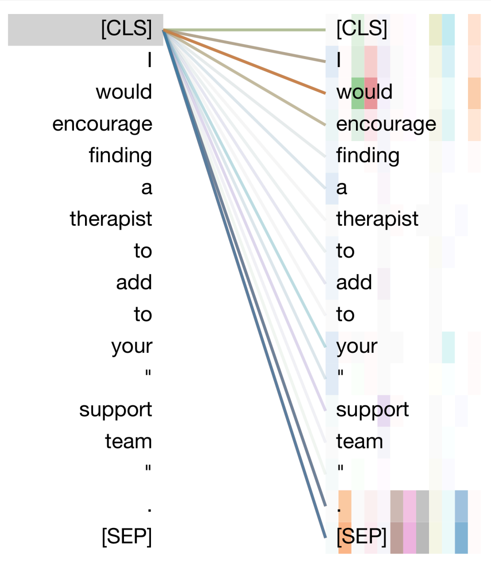
Flair results on test set.

ATTENTION



Attention weights visualized using BertViz (Vig 2019)

ATTENTION



Attention weights visualized using BertViz (Vig 2019)