Tweet Classification to Assist Human Moderation for Suicide Prevention

Ramit Sawhney (ramits.co@nsit.net.in), Harshit Joshi, Alicia Nobles, Rajiv Ratn Shah

Research Questions

Are there temporal variations in linguistic features that differentiate tweets containing expressions of suicidal intent that could be misidentified as suicidal intent (i.e., edge cases)?

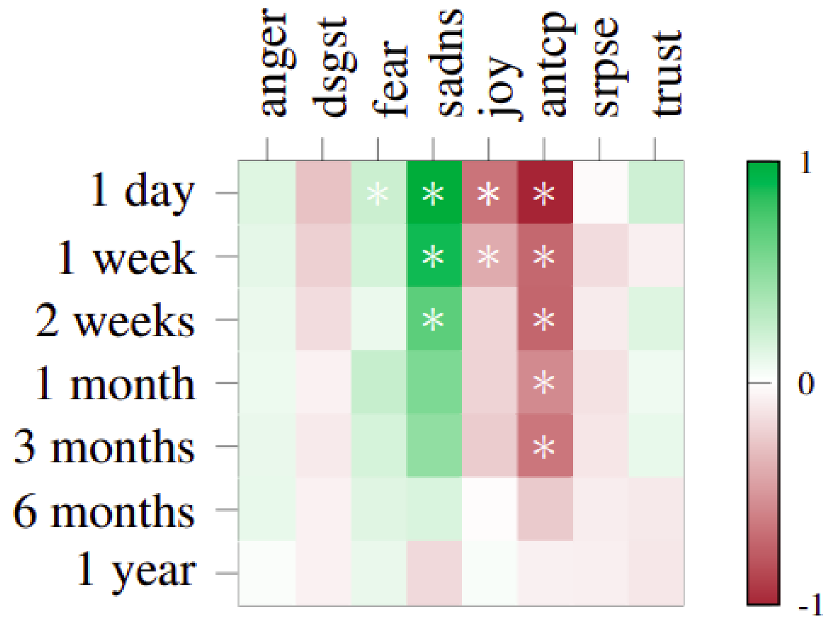
Using a Sparse Additive Generative Model (SAGE), we analyze how a user's language in their tweets varies temporally.

Suicidal Intent Present		Suicidal Intent Absent	
1 Day	**SAGE**	**1 Day**	**SAGE**
slit	2.91	dispatch	2.48
needles	2.78	neverland	2.44
schizophrenia	2.42	runaways	2.16
antidepressants	2.23	lobbying	2.11
urges	2.13	shutdown	2.05
1 Week	**SAGE**	**1 Week**	**SAGE**
selfloathing	2.78	bandersnatch	2.84
symbols	2.14	braveheart	2.77
resigned	2.13	birdbox	2.68
miscarriage	1.98	copycat	2.39
storytelling	1.71	lmmfaooooo	2.31
2 Weeks	**SAGE**	**2 Weeks**	**SAGE**
cbd	2.38	hamper	2.00
merry	1.46	camels	2.00
hearts	1.56	glances	1.90
pharma	1.26	obscene	1.90
reflux	1.12	reindeer	1.88

Five cherry-picked distinctive words across time‐buckets obtained using SAGE for historic tweets prior to the tweet in question. A higher SAGE score is indicative of its saliency.

Are there temporal variations in emotional language that differentiate between tweets containing expressions of suicidal intent and edge cases?

We fine-tune a pre-trained LM for emotions and use it to automatically extract the differentiating temporal variations in emotions expressed in tweets.



Temporal variation in the eight primary emotions expressed for historical tweets prior to the tweet in question (here we visualize a tweet where SI is present)

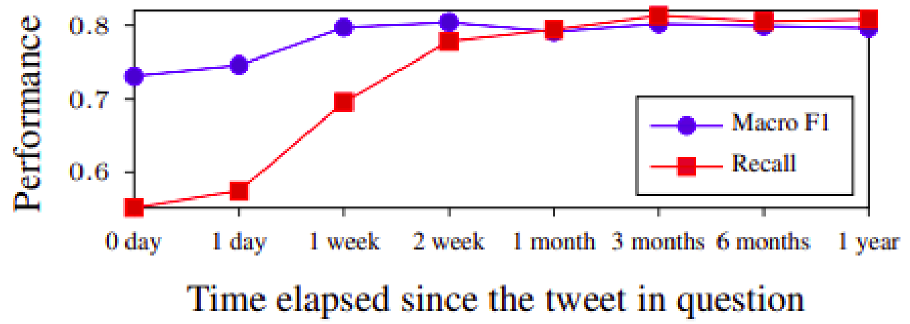
Can predictive models, trained jointly on temporal activity and language features, differentiate between tweets containing expressions of suicidal intent and edge cases?

We build a time-aware sequential neural model that differentiates between tweets.

We then examine the interpretability of the model's decision on three example tweets.

Model	Macro F1 ↑	Precision ↑	Recall ↑
Random Forest	0.536	0.489	0.513
Logistic Regression	0.571	0.563	0.583
C-LSTM	0.588	0.568	0.597
SDM	0.743	0.578	0.755
DualContextBert	0.767	0.589	0.786
Exponential Decay	0.737	0.582	0.759
Surprise and Episodic Modeling	0.741	0.583	0.762
STATENet + Temporal Attention	**0.804***	**0.612***	**0.813***

How much historical context is useful?



Acknowledgement
Dr. Nobles was supported by NIH NIDA K25 DA049944
We thank the anonymous reviewers for their valuable input.