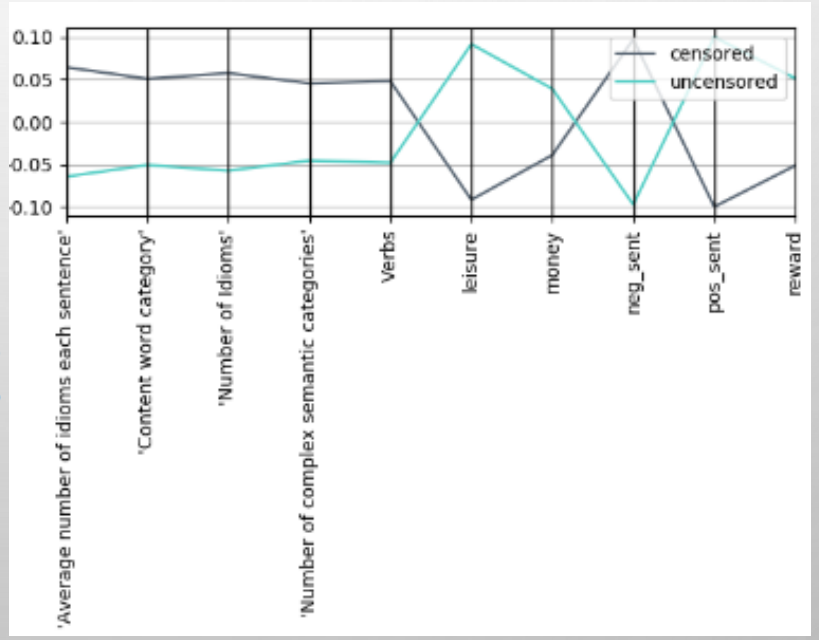
NEURAL NETWORK PREDICTION OF CENSORABLE LANGUAGE

CHRIS LEBERKNIGHT, ANNA FELDMAN, MONTCLAIR STATE UNIVERSITY, NSF 1704113

- **Research Highlight** : We find consistent linguistic features that seem to characterize censored and uncensored texts independently.
- 4 sets of linguistic features from 2 datasets – LIWC features, CRIE features, semantics features, and the number of followers feature.
- Our model predicts censored tweets using only linguistic features performs with a 88.50 % accuracy

dataset	N	H	features	accuracy
baseline				49.98
human baseline (Ng et al., 2018b)				63.51
scraped	500	50,50,50	Seed 1	80.36
scraped	800	60,60,60	Seed 1	80.2
Zhu et al's	800	50,7	Seed 1	87.63
Zhu et al's	800	30,30	Seed 1	86.18
both	800	60,60,60	Seed 1	75.4
both	500	50,50,50	Seed 1	73.94
scraped	800	30,30,30	all except LIWC	72.95
Zhu et al's	800	60,60,60	all except LIWC	70.64
both	500	40,40,40	all except LIWC	84.67
both	800	20,20,20	all except LIWC	88.50
both	800	30,30,30	all except LIWC	87.04
both	800	50,50,50	all except LIWC	87.24

High performance obtained excluding the LIWC features shows that the key to distinguishing between censored and uncensored posts seems to be the features related to writing style, readability, sentiment, and semantic complexity of a text.



Best features that contribute to distinguishing censored and uncensored posts suggest that the censored posts generally convey more negative sentiment and are more idiomatic and semantically complex in terms of word usage.