Tracing Linguistic Markers of Influence in a Large Online Organisation

Prashant Khare*, Ravi Shekhar†, Vanja Mladen Karan*, Stephen McQuistin‡, Colin Perkins‡, Ignacio Castro*, Gareth Tyson*§, Patrick G.T. Healey*, Matthew Purver*¶

*Queen Mary University of London, †University of Essex, ‡University of Glasgow
§Hong Kong University of Science & Technology, ¶Jožef Stefan Institute
{p.khare, m.karan, i.castro, g.tyson, p.healey, m.purver}@qmul.ac.uk,
r.shekhar@essex.ac.uk, sm@smcquistin.uk, csp@csperkins.org

Abstract

Social science and psycholinguistic research have shown that power and status affect how people use language in a range of domains. Here, we investigate a similar question in a large, distributed, consensus-driven community – the Internet Engineering Task Force (IETF), a collaborative organisation that develops technical standards for the Internet. Our analysis, based on lexical categories (LIWC) and BERT, shows that participants' levels of influence can be predicted from their email text, and identifies key linguistic differences (e.g., certain LIWC categories, such as WE are positively correlated with high-influence). We also identify the differences in language use for the same person before and after becoming influential [1].

1 Introduction and Related Work

Motivation Online communities are rapidly growing. It is imperative to study them to gain a better understanding of online dynamics and important processes such as decision-making. Prior work has shown that influence is an important aspect to consider while analysing online community dynamics (Bapna and Umyarov, 2015; Vega et al., 2021). Social and psycholinguistic research has also revealed that a person's power and status (i.e., influence) is reflected in their usage of language (Nguyen et al., 2016; Guinote, 2017). In this paper, we focus on linguistic traits exhibited by influential people in a large online community.

Detecting meaningful domain-independent indicators of influence is difficult (Danescu-Niculescu-Mizil et al., 2012). Instead, we focus on the Internet Engineering Task Force[2] (IETF) - a large, open, voluntary, standards developing organisation with over 2M emails between 56k participants over 20 years. The decentralised, consensus-oriented nature of the IETF makes it an interesting case study for two reasons. First, compared to the social media data commonly used in similar studies (e.g. Tchokni et al., 2014; Prabhakaran, 2015), IETF emails are usually longer and goal-oriented. Second, the IETF is a decentralised organisation where the decision-making is collaborative and consensus-driven (Bradner, 1996; Resnick, 2014). Hence, the resulting social interactions are very different to alternative email-based datasets such as the Enron Corpus (Klimt and Yang, 2004), or interactions with more rigidly defined power distinctions e.g., admins/users, judges/lawyers (Danescu-Niculescu-Mizil et al., 2012).

Related Work Most studies of influence either focus on community structure rather than language, or use language indirectly. Urena et al. (2019) give a survey of the former approach. In an example of the latter, Prabhakaran et al. (2014) compare users with different influence in terms of their linguistic similarity or *co-adaptation*, the increasing similarity of interlocutors to each other in how they use language (see also Danescu-Niculescu-Mizil et al., 2012; Ver Steeg and Galstyan, 2013; Noble and Fernández, 2015; Kawabata et al., 2016; Buske, 2019; Healey et al., 2023). Some studies (Bramsen et al., 2011; Gilbert, 2012) do focus on modelling influence from text of Enron emails by identifying keywords/phrases that indicate influence. Rosenthal (2014) and Tchokni et al. (2014) extend this approach to other domains, including Twitter, Wikipedia talk pages, and debates, and include a wider range of linguistic markers.

Goals We focus on discovering linguistic markers of influence in a large consensus-driven standards developing organisation, where the consensus is based on elaborate discussions between participants on mailing lists. To complement this analysis, we also study the linguistic behaviour

[1] Code: https://github.com/sodestream/acl2023-tracing-linguistic-markers

[2] IETF is responsible for producing technical standards for internet infrastructure. https://www.ietf.org/

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 82–90
July 9-14, 2023

of participants at different hierarchical levels in IETF, as well as participants in different periods of their participation, similar to Danescu-Niculescu-Mizil et al. (2013), who considered the behaviour of participants as a measure of influence and claim that participants tend to echo the linguistic style of influential individuals. We map this to three research questions: **RQ1:** *How do linguistic traits differ between more and less influential participants?* **RQ2:** *How do linguistic traits vary for participants at different levels of the organisation hierarchy?* **RQ3:** *How does linguistic behaviour of participants change as they gain influence?*

2 Methodology

We aim to understand the correlation between influence, as defined by either network-based centrality metrics (*mail-based*) or organisational role influence (*role-based*), and language usage in terms of linguistic traits. For each participant, we consider the emails they sent in a given time period and investigate correlations of certain features of their email text with two different measures of influence.

LIWC Representation Linguistic Inquiry and Word Count (LIWC, Pennebaker et al., 2015) is a well-recognised psycholinguistic lexicon; it provides word counts for 85 different linguistic, psychological, personal concern, and informal language marker categories. Here, we aggregate the word counts within each linguistic category for each participant using the LIWC 2015 dictionary (academic license), and normalise by the total number of emails sent by that participant. Such a normalisation is more appropriate here than normalising by total number of words written, as many IETF emails include long technical sections. This generates a representation of a participant as their mean usage of each LIWC category; while this is a relatively reduced, low-dimensional representation of a person's language, it has the advantage of being interpretable and psychologically well-motivated.

BERT Representation The LIWC representation ignores context. To allow comparison to more advanced methods, we use the context-dependent representations from BERT (Devlin et al., 2019) via the open-source HuggingFace library (Wolf et al., 2019). The participant-specific BERT representation is calculated by averaging the text representations (last layer *CLS* vectors) over all their emails.

3 Experimental Set-up

Dataset The IETF is organised in Working Groups (WGs). Each WG has a technical focus (e.g., HTTP WG for the HTTP protocol) and one or more WG chairs. We use data from two public sources: the IETF mail archives[3] and the Datatracker[4]. The mail archives cover WG activities, meetings, and administration. We gathered 2,106,804 emails from 56,733 email addresses spanning 2000-2019.

To determine *mail-based* influence, we use a social graph based on mailing list interactions (messages from one person to another) as built by Khare et al. (2022). We rank participants by their eigenvector centrality, a measure of a node's influence in a graph, and transform rank to a percentile. To determine *role-based* influence, we used Datatracker for information about WG chairs and their tenure.

RQ1 (mail-based influence) We used a 5-year subset of the data for RQ1 due to the computation cost, still giving a reasonable period to observe the participation consistency in the IETF community (McQuistin et al., 2021; Khare et al., 2022). We took data from 2015-2019 with 300,806 emails from 5,363 unique participants. This subset has 212,253 unique tokens, as opposed to 735,605 unique tokens in the whole dataset, and the median length of emails is 504. We calculate the *mail-based* influence score and LIWC representation[5] for each participant as described. We fit a linear regression model using LIWC representations to predict influence percentile and observe the magnitude and directions of significant coefficients.

RQ2 (role-based influence) While *mail-based* influence was crucial to consider the activities of the participants based on the email network, *role-based* influence is equally crucial as they are involved in organisational decision making.[6] We use the same time period as in RQ1, but here we predict organisational *role-based* influence. We split the data into two categories: (a) WG chairs and (b) participants who have never been WG chair. We

[3] https://mailarchive.ietf.org/

[4] https://datatracker.ietf.org/ – the administrative database of the IETF, containing metadata about participants and their roles, working groups, document status, etc.

[5] We filter out 104 ambiguous words that are present in LIWC but have technology, security, and network context meaning in IETF, using manually curated lists, for e.g., attack, argument, secure etc. We do this across all RQs.

[6] In the top 10% *mail-based* influential participants, less than 30% are WG chairs with significant *role-based* influence.

calculate the LIWC representations for each person, train a logistic regression model to predict category, and observe the LIWC category coefficients.

RQ3 (changes in influence) We look at participants who went from low to high influence over time: individuals who had a *mail-based* influence below the 50th percentile when they joined the IETF, and reached the top 10th percentile at some point. For each participant, we generate two different representations based on two periods — the year of joining and year of reaching the top 10th percentile for the first time — and assign these to two different classes. As in RQ2, we then train a logistic regression model to predict these classes, and examine the coefficients of the LIWC categories.

BERT-based variants Our primary purpose is not to assess the predictive power of LIWC representations, but to use them as a tool to characterise linguistic variations in a meaningful way. However, in order to understand their predictive potential, given their relatively simple nature, we compare them to BERT. For these comparisons, we use the BERT representations described in Section 2.

For each RQ we use the same experimental setup as described above. We split the data 80:20 into train and test set and train a prediction model (regression for RQ1 and classification for RQ2 & RQ3). To experiment with both linear and non-linear models, we include linear and logistic regression and multi layer perceptrons, using implementations from scikit-learn (Pedregosa et al., 2011) with default parameters. As evaluation metrics we used Pearson's ρ and macro-F1 score.

4 Results & Discussion

We now explore the results (see Table 1 for all experiments) and answer our research questions.

4.1 Answers to RQs

RQ1 — The following LIWC categories are significantly correlated ($p < 0.05$) with higher *mail-based* influence: WE, INFORMAL, RISK, ADJECTIVE, ANGER, THEY, and BIO. Categories such as NETSPEAK, SEXUAL, HEALTH, DEATH, BODY are correlated with lower influence. This suggests that influential people tend to indicate a collaborative and community-oriented approach with first-person plural (WE) and third-person plural category (THEY) usage. This is consistent with Kacewicz et al. (2014) and Guinote (2017), who show that influential people use more first-person plural. They also use more organisational language, which is shown by the negative correlation of informal slang language categories (NETSPEAK, SEXUAL, BODY). We see some unexpected hidden trends due to word ambiguity (e.g., words like '*trust*' and '*live*'), which are investigated in Section 4.2.

RQ2 — From 1, we see that working group (WG) chairs are more social and collaborative, as is shown by WE and SOCIAL categories. This is in line with similar findings from RQ1 and also about leadership engagements from previous works (Strzalkowski et al., 2012; Liu, 2022; Kacewicz et al., 2014; Guinote, 2017; Akstinaite et al., 2020). Also, WG chairs use tentative statements (TENTAT) in discussions, primarily focused on technical feedback and revisions, or suggesting alternatives. Examples showcasing the use of words such as *'or'* and *'seems'*-

- 'seems': *"With the risk of disturbing with statements, but avoiding too many questions:This seems against the goal of reducing headers."*
- 'or': *"Question is do we need to carry around an outer IP-in-IP header for that or not?"*

RQ3 — From Table 1, we observe that when participants become *mail-based* influential they are likely to be more descriptive and engaged in immediate state of issues and situations as seen from the correlation of auxiliary verbs (AUXVERB), adverb, risk, and present focus (FOCUSPRESENT). They are also more involved in cognitive processes (COGPROC) as compared to their previous self when they were new to IETF and had little influence.

4.2 Discussion

To better understand these LIWC categories and what kind of words play a role in the behaviour of individual categories, we calculate the frequency of words in each LIWC category as they appear in the emails. Next, we consider the top 30 most frequent words in each LIWC category and perform regression analysis on *mail-based* influence for participants, but using only these 30 words as features to generate the participant representation. We conducted this experiment separately for each LIWC category that was significant in the first experiment.

From the word based analysis we make multiple observations. E.g., words like *'we'* imply a collective approach and is strongly correlated with the higher influence. Similarly, the use of word *'well'*

RQ1	High influence	BIO, WE, INFORMAL, THEY, NEGEMO, ANGER, RISK, ADJECTIVE
	Low influence	SEXUAL, DEATH, INGEST, NETSPEAK, HEALTH, FEMALE, BODY, AFFILIATION, CONJ
RQ2	WG Chair influence	TENTAT, IPRON, SOCIAL, SEE, FEEL, WE
	non-WG Chair	COGPROC, RELATIV, AFFILIATION, I, REWARD
RQ3	Top 10 percentile	ADVERB, PREP, ANGER, AUXVERB, MALE, COGPROC, ACHIEV, RISK, FOCUSPRESENT
	Below 50^{th} percentile	FUNCTION, PPRON, SHEHE, IPRON, NUMBER, CERTAIN, SEXUAL, INFORMAL

Table 1: LIWC categories where $p < 0.05$.

is standard, such as politely resuming the conversation (e.g., '*well, I agree*') or providing an approval over something (e.g., '*this works as well*'). These words are well associated with the influential participants. Otherwise, influential participants are generally not observed to be informal and other frequent words (other than *'well'*) within INFORMAL category do not demonstrate a strong correlation with the growing influence. Also, *'well'* is the most frequent word in the INFORMAL category.

More influential people (both *mail-based* and *role-based*) are also observed to engage more in IETF communities. The conversations can often reflect situations where, as a part of review and feedback process, more influential people highlight limitations in protocol standards, stress on specifics, and compare with existing protocols or previous versions. Several words across different LIWC categories (RISK, NEGEMO, and ADJ) highlight such behaviour, e.g., *'problems'*, *'before'*, *'particular'*, *'specific'*, *'different'*, *'most'*, and *'than'*.

However, there are many words with dual sense, like *'trust'* which has a very technology specific usage related to network security instead of conversations involving trust issues between individuals or trust in any given situation. Similarly, the word *'live'* is related with an application or network being live, instead of its conventional meaning. We also observed that some of the LIWC categories, such as BIO, did not have specific terms that could clearly establish its significance in favour of influential participants (e.g., word *'problems'* and *'trust'* reflecting the significance for the category RISK), instead such categories had several words with quite weak correlation with influential participants. Such words collectively drifted the weight of the category towards influential participants.

4.3 BERT-based results

We compared the performance of the LIWC- and BERT-based models. Results in Table 2 indicate our LIWC approach is better than an intuitive BERT-based baseline. We hypothesize that the reason for this is that LIWC is specialised to detect linguistic markers relevant for this task. Also, to ensure fair comparison, BERT representations were not fine-tuned for the tasks. We believe combining LIWC and BERT might give better representations, especially when dealing with ambiguous words. Curiously, when observing t-SNE (Van der Maaten and Hinton, 2008) projections of participants' BERT representations (Appendix A), we find that low-influence users show a much bigger variation for relevant categories such as WE, NETSPEAK and INFORMAL. We will investigate this in future.

	LIWC		BERT	
	LR	MLP	LR	MLP
RQ1 (Pearson ρ)	0.850^*	0.852^*	-0.018	0.015
RQ2 (Micro F_1)	91.21	92.46	87.69	92.21
RQ3 (Micro F_1)	88.89	90.74	51.85	55.56

Table 2: LIWC vs BERT(* $p < 0.0001$.)

5 Conclusions & Future Directions

This paper explores the linguistic patterns of influence in an online collaborative organisation, by analysing the differences between high- and low-influence participants. Using two aspects of influence — *mail-based*, derived from the email network, and organisational *role-based* — we were able to unfold several traits that differentiate influential participants from others. Many of our findings seem corroborated by studies in organisational theory. We observed that influential people exhibit more collaborative and community-oriented traits, and also stronger signs of engagement in discussions. We also observed that as people go on to become influential participants, they evolve in their communication and are seen to be more engaging and descriptive in their linguistic style. An interesting practical application of our research is identifying and analyzing groups that are dysfunctional in terms of participant roles and their communication patterns (e.g., where the chair is not performing their role). In future work, we will

extend the experiments to study these patterns of interaction in more linguistic depth, between more different roles within an organisation (possibly for multiple collaborative organisations). We will attempt to go beyond lexical count and account for word context.

6 Limitations

One of the main limitations is that we used the standard LIWC-based analysis approach, which is purely lexical and does not take into account the context in which a word appears. Consequently, many words that have very specific senses in the context of the IETF get miscounted as occurrences of LIWC categories. This could be addressed by a more advanced method of mapping to LIWC categories that would account for context. Another limitation is that we manually generated a filtering list containing words specific to the IETF. This list might not be exhaustive enough. Also, we were limited by not conducting an exhaustive hyper-parameter search on our models. We also understand that many emails are longer than 512 tokens (the input limit of the BERT model we used) and might have not been captured completely by our BERT model. However, most of the emails do fit into this BERT sequence length limit. We did not fine tune BERT on the IETF data; this might have given better performance, although it is not clear if it would have given more insight: our main goal is not performance but analyzing/comparing characteristics of existing models. It is also worth highlighting that the data used in this work is strictly in English, and the psycholinguistic categories in LIWC are also based on English language. Hence, this study may be biased and not fully capture variations in linguistic traits that are culturally agnostic.

Ethical considerations — Participation in the IETF is bound by agreements and policies explicitly stating that mailing list discussions and Datatracker metadata will be made publicly available.[7] We use only this publicly available data in our analysis. We have discussed our work with the IETF leadership to confirm that it fits their acceptable use policies. We have also made provisions to manage the data securely, and retain it only as necessary for our work.

[7] See both https://www.ietf.org/about/note-well/ and the IETF privacy policy available at https://www.ietf.org/privacy-statement/.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was supported by the UK EPSRC under grants EP/S033564/1 and EP/S036075/1 (Sodestream: Streamlining Social Decision Making for Enhanced Internet Standards). Purver was also supported by the Slovenian Research Agency via research core funding for the programme Knowledge Technologies (P2-0103).

References

Vita Akstinaite, Graham Robinson, and Eugene Sadler-Smith. 2020. Linguistic markers of ceo hubris. *Journal of Business Ethics*, 167:687–705.

Ravi Bapna and Akhmed Umyarov. 2015. Do your online friends make you pay? a randomized field experiment on peer influence in online social networks. *Management Science*, 61(8):1902–1920.

Scott O. Bradner. 1996. The Internet Standards Process – Revision 3. RFC 2026.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782.

Jakob A Buske. 2019. Linguistic accommodation between leaders and followers. B.S. thesis, University of Twente.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1037–1046.

Ana Guinote. 2017. How power affects people: Activating, wanting and goal seeking. *Annual review of psychology*, 68:353–381.

Patrick Healey, Prashant Khare, Ignacio Castro, Gareth Tyson, Mladen Karan, Ravi Shekhar, Stephen McQuistin, Colin Perkins, and Matthew Purver. 2023. Power and vulnerability: Managing sensitive language in organisational communication (extended abstract). In *ST&D 2023: Annual Meeting of the Society for Text and Discourse*, June 28 – June 30, 2023, Oslo, Norway.

Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.

Kan Kawabata, Visar Berisha, Anna Scaglione, and Amy LaCross. 2016. A convex model for linguistic influence in group conversations. In *INTERSPEECH*, pages 1442–1446.

Prashant Khare, Mladen Karan, Stephen McQuistin, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, and Ignacio Castro. 2022. The web we weave: Untangling the social graph of the IETF. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 500–511.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

Amy H Liu. 2022. Pronoun usage as a measure of power personalization: A general theory with evidence from the chinese-speaking world. *British Journal of Political Science*, 52(3):1258–1275.

Stephen McQuistin, Mladen Karan, Prashant Khare, Colin Perkins, Gareth Tyson, Matthew Purver, Patrick Healey, Waleed Iqbal, Junaid Qadir, and Ignacio Castro. 2021. Characterising the IETF through the lens of RFC deployment. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 137–149.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Bill Noble and Raquel Fernández. 2015. Centre stage: How social network position shapes linguistic coordination. In *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*, pages 29–38.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report.

Vinodkumar Prabhakaran. 2015. *Social power in interactions: Computational analysis and detection of power relations*. Ph.D. thesis, Columbia University.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486.

Pete Resnick. 2014. On Consensus and Humming in the IETF. RFC 7282.

Sara Rosenthal. 2014. Detecting influencers in social media discussions. *XRDS: Crossroads, The ACM Magazine for Students*, 21(1):40–45.

Tomek Strzalkowski, Samira Shaikh, Ting Liu, George Aaron Broadwell, Jenny Stromer-Galley, Sarah Taylor, Umit Boz, Veena Ravishankar, and Xiaoai Ren. 2012. Modeling leadership and influence in multi-party online discourse. In *Proceedings of COLING 2012*, pages 2535–2552.

Simo Editha Tchokni, Diarmuid O Séaghdha, and Daniele Quercia. 2014. Emoticons and phrases: Status symbols in social media. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Raquel Urena, Gang Kou, Yucheng Dong, Francisco Chiclana, and Enrique Herrera-Viedma. 2019. A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. *Information Sciences*, 478:461–475.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Lea Vega, Andres Mendez-Vazquez, and Armando López-Cuevas. 2021. Probabilistic reasoning system for social influence analysis in online social networks. *Social Network Analysis and Mining*, 11(1):1–20.

Greg Ver Steeg and Aram Galstyan. 2013. Information-theoretic measures of influence based on content dynamics. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 3–12.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Appendix A: BERT-based results

We investigated how BERT representations vary for participants, as per influence, across different significant LIWC categories. For each participant, we calculated the LIWC category representation by averaging the BERT representation of the words in that LIWC category and then projected using t-SNE. As Figures 1, 2 and 3 show, high-influence participants show less variation in their BERT representations compared to lower-influence participants, for the LIWC categories WE, NETSPEAK and INFORMAL respectively.

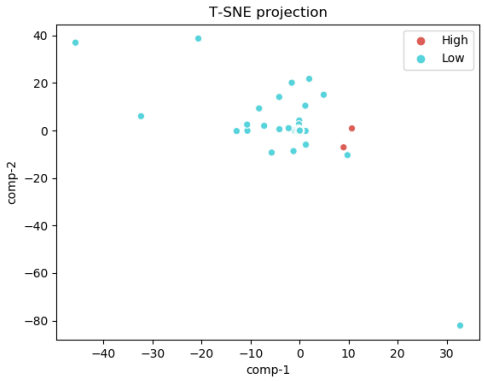

Figure 1: WE category representation

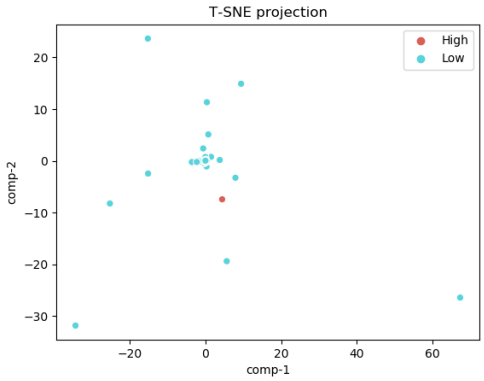

Figure 2: NETSPEAK category representation

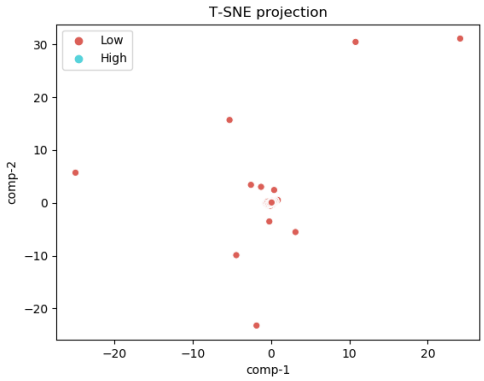

Figure 3: INFORMAL category representation

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 6 in Limitations section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 2

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2 - we used artifact(s) as they they were intended to without any modifications.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We have used a publicly available dataset as allowed by IETF's privacy statement https://www.ietf.org/privacy-statement/

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2 LIWC Representation

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Section 3

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We used default parameters for experiments without parameter tuning.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3 (default parameters)

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2 and Section 3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.