Seasonality Visualizations of Online Text

Andrea W Wang*, Allison Koenecke, David Mimno
(*aww66@cornell.edu)

Cornell Bowers CIS Information Science

Introduction

- Existing work studying languages of online communities often lack considerations for the diversity of temporal linguistic patterns. In this work, we contribute a methodological guide for practitioners to visualize seasonality in online text data, and provide case studies to showcase the range of temporal linguistic patterns found across subreddits.
- **Research question: How to visualize temporal linguistic patterns?**

Motivating Examples

Comments most similar to **"I was at Mt. Baker WA yesterday and it was just as good. No clear skies though."** *(r/snowboarding, 2018 Jan)* tend to be in winter.

- That looked like a great day, perfect weather, plenty of virgin snow and nice powder. Also the view ain't bad. Do you go to Mt Baker often? *(2015 Oct)*
- This was yesterday, but the conditions were great. A little windy, but I can't really complain. *(2014 Feb)*
- I just came back from Whistler and did the same thing. They said Friday was one of the best days with little to no clouds or wind. Hope you went that same day. *(2018 Mar)*

Methodology

- **Data** include comments from 2014 Jan to 2018 Oct in 84 subreddits. For each subreddit, we randomly sampled 500,000 comments with more than 5 words.
- **Similarity Measure "EMB"**: We use **Sentence-T5** to generate comment embeddings, then we average comment embeddings in a given month to get month embedding. Lastly, we get similarity measure, "EMB", by calculating Euclidean distance between two month embeddings.

Conclusion

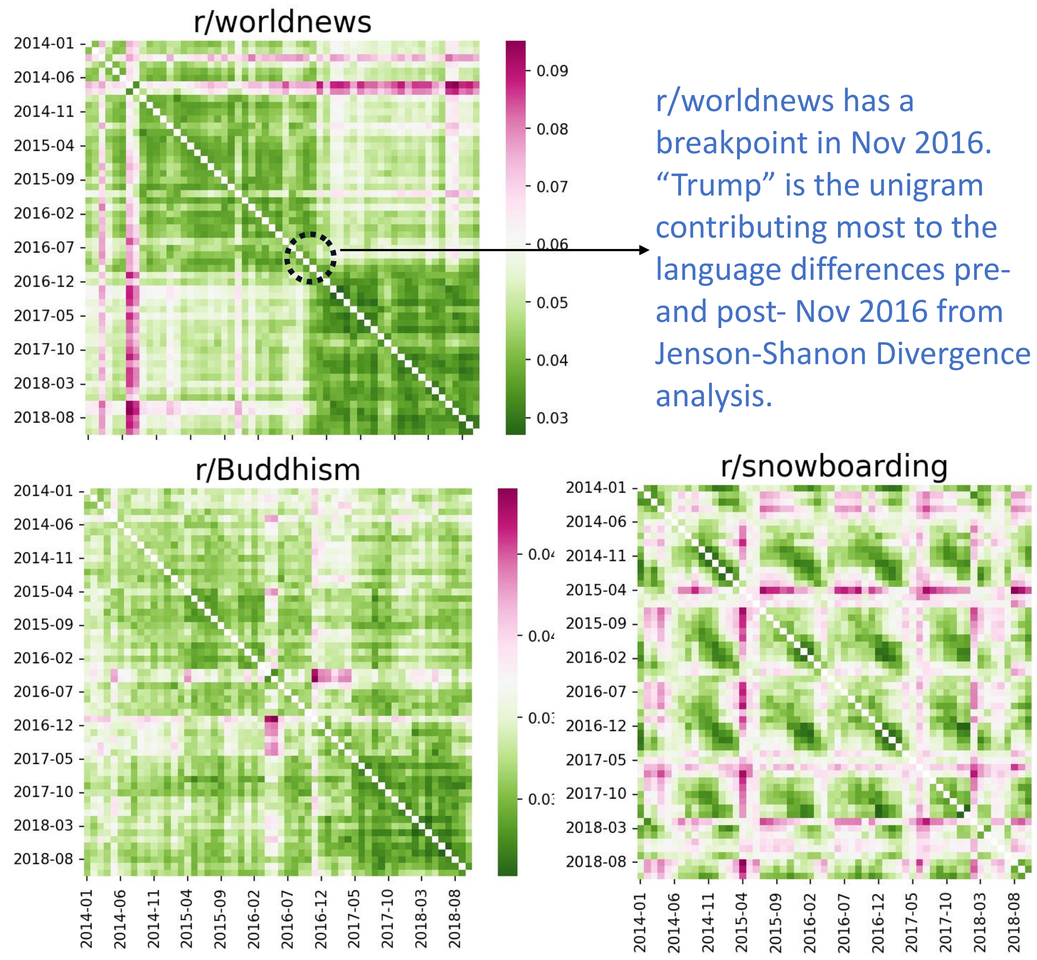
- Through visualization, we find diverse temporal linguistic patterns in online communities, from transient to seasonal.
- We recommend choosing visualization methods according to usage:
  - •Method (1) to study **patterns tied to specific months of interest**,
  - •Method (2) to show **general temporal linguistic patterns** where specific time effect is not of interest,
  - •Method (3) to **compare across larger number of online communities.**
- We encourage practitioners to use these simple visualizations to consider seasonality when applying NLP methods to temporal data.

References

- Jianmo Ni et al,. 2022. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.
- Jonathan P. Chang et al., 2020. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.
- Justine Zhang et al. "Community identity and user engagement in a multi-community landscape". In: Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. 1. 2017, pp. 377–386.
- Cristian Danescu-Niculescu-Mizil et al,. 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 (WWW '13). Association for Computing Machinery, New York, NY, USA, 307–318.

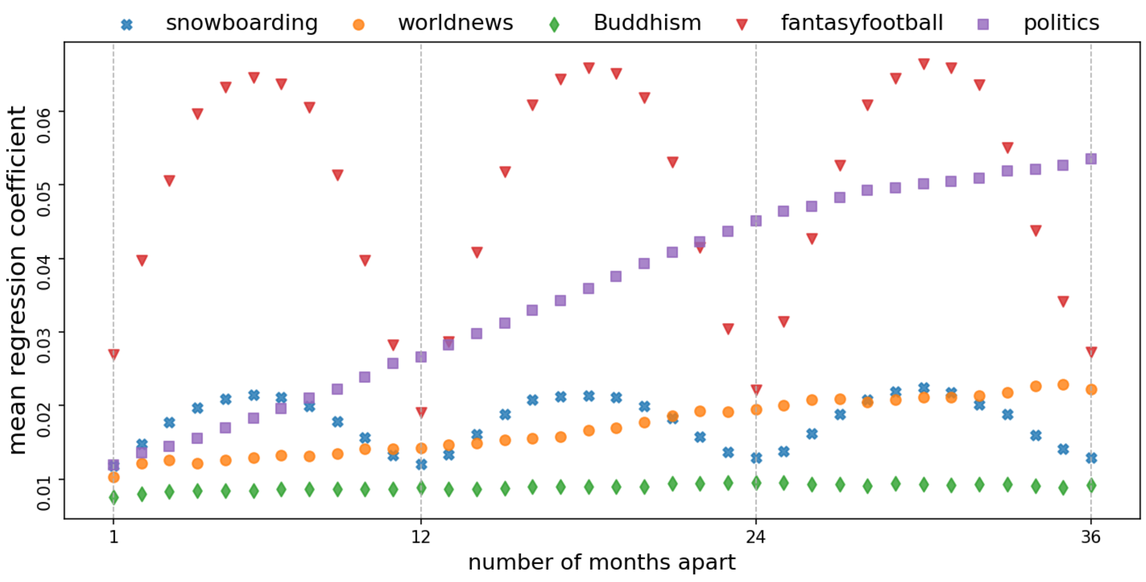
Three Visualization Methods

METHOD 1. Pairwise EMB with heatmaps.



r/worldnews has a breakpoint in Nov 2016. "Trump" is the unigram contributing most to the language differences pre- and post- Nov 2016 from Jenson-Shanon Divergence analysis.

METHOD 2. Coefficients of num_months apart from OLS regression: EMB ~ C(num_months_apart) + time_fixed_effect between weeks.



METHOD 3. Coefficients of is_same_month & is_same_year from OLS regression: EMB ~ C(is_same_month) + C(is_same_year) between weeks.