9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Seasonality Visualizations of Online Text

Keywords: computational linguistics, text as data, seasonality, time series, visualization

Extended Abstract

Computational linguistics techniques are regularly used to study socialization in online communities through user engagement patterns [3]. However, the existing work focusing on community language often avoids accounting for seasonality by only considering a short time frame of online comments. Meanwhile, the existing work that *does* consider seasonality in online communities seldom uses NLP tools that go beyond word-count-based methodology. In this paper, we bridge this gap by performing a large-scale study of seasonality in subreddit linguistic patterns using a word-sequence-based model. We contribute a methodological guide for practitioners to visualize seasonality in online text data, and provide three case studies showcasing the range of seasonality found across subreddits.

Data We used a Reddit corpus from `ConvoKit` [1] to extract comments from 84 subreddits between January 2014 to October 2018. The subreddits we studied spanned diverse topics such as sports, business, religions, special interests, and daily advice. For each subreddit, we randomly sampled 500,000 comments each containing more than 5 words. To illustrate the diversity of seasonal linguistic patterns, we present *r/snowboarding*, *r/worldnews*, and *r/Buddhism* as case studies.

Metric Methodology To explore month-to-month linguistic similarity, we utilized a word-sequence-based model. Specifically, we generated a vectorized representation ("comment embedding") for each comment using the Sentence-T5 encoder from a pre-trained text-to-text transformer, T5 [2]. Subsequently, we calculated a vectorized representation ("month embedding") for each month in a subreddit by averaging the comment embeddings corresponding to that month. Finally, we determined the Euclidean distance between the month embeddings, producing our metric, referred to as "EMB." A lower EMB distance corresponds to higher temporal similarity.

Seasonality Visualizations In Fig 1, we present **heatmaps** for our three case studies, illustrating the pairwise EMB between each pair of months from January 2014 to October 2018. Recall that EMB is a distance metric: green cells indicate linguistic similarity and red cells indicate linguistic dissimilarity between monthly pairs. Visually tracing the corresponding grid patterns can aid seasonality interpretations in different settings. For example, the granular checkered pattern in *r/snowboarding* represents a strong linguistic similarity every winter with other winters, and less similarity between summers, likely due to the nature of the sport. Meanwhile, *r/worldnews* reveals a two-by-two checkered pattern, with major seasonality differences pre- and post- November 2016. Using Jenson-Shanon Divergence analysis, we found that "Trump" is the most significant unigram contributing to the temporal differences (Fig 2). Finally, *r/Buddhism* is a more temporally stable community with a minimal grid pattern, corresponding to minimal differences in linguistic similarity over time.

Using the visualization methodology presented in Fig 1, it is straightforward to understand seasonal trends tied to specific months of interest between 2014 and 2018. However, the heatmap is high dimensional and thus less user-friendly. Furthermore, it doesn't consider granularity of time trends within-month, and may be difficult to decipher less-seasonal patterns such

as those in *r/Buddhism*. We can ameliorate these concerns by introducing a second visualization of textual seasonality that both reduces dimensionality of plots and captures within-month trends, though it loses the visualization of specific months between 2014 and 2018.

In Fig 3, we plot **coefficient-based** visualizations for each of *r/snowboarding*, *r/worldnews*, and *r/Buddhism*. Using recalculated weekly-level embeddings per our metric methodology, we now run an ordinary least-squares regression with time fixed effects, and estimate the EMB distance between each week pair for a single subreddit.[1] Our coefficient of interest is the categorical variable *number of months apart* between pairs of weekly embeddings, plotted along the x-axis of Fig 3; recall that a lower coefficient (y-axis) corresponds to a lower EMB distance (i.e., closer similarity). We see the clearest seasonal pattern in *r/snowboarding* where comments in the same month are consistently most linguistically similar, regardless of year (i.e., *number of months apart* is a multiple of 12). On the other hand, *r/worldnews* showed a significant increasing trend over time, indicating that a larger temporal distance corresponds to a larger linguistic distance (and more dissimilar comments). Lastly, *r/Buddhism* coefficients are more stable over time with only a slight positive trend.

Finally, we extend the scope of our visualizations outside of our three case study subreddits, and comment on how to generate visualizations of seasonality across multiple subreddits. Using a similar **regression-based** methodology on normalized EMB distances, we now additionally focus on two covariates: a binary variable indicating whether two EMB weeks are in the same year, and a binary variable indicating whether two EMB weeks are in the same month. Plotting the coefficients for these two covariates as the x- and y-axes of Fig 4, we can identify which of our 84 subreddits have certain trends. The lower-right quadrant shows a set of outliers representing sports subreddits, which are extremely seasonal (showing low EMB distances during the same months of each year, and high EMB distances across years). In contrast, subreddits like *r/AmITheAsshole* and *books* do not see large seasonal differences in comments.

Discussion Our work contributes to the study of online communities by providing three visualization methods for practitioners to understand and interpret the seasonality of text data in online communities. We encourage practitioners to consider seasonality when applying NLP methods to temporal data. This work has implications in the assumptions made in user behavior: e.g., new users joining in different times of year may be fundamentally different, and community activity during some months may disproportionately influence new users' socialization patterns (e.g. adopting community language).

References

[1] Jonathan P. Chang et al. "ConvoKit: A Toolkit for the Analysis of Conversations". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2020, pp. 57–60. URL: https://aclanthology.org/2020.sigdial-1.8.

[2] Jianmo Ni et al., eds. *Sentence-T5: Scaling up Sentence Encoder from Pre-trained Text-to-Text Transfer Transformer*. 2022. URL: https://aclanthology.org/2022.findings-acl.146/.

[3] 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.

[1] We ran 500 regressions bootstraping over $n = 50,000$ samples for each subreddit; p-values of all the variables in three case studies were significant in every bootstrapped regression.

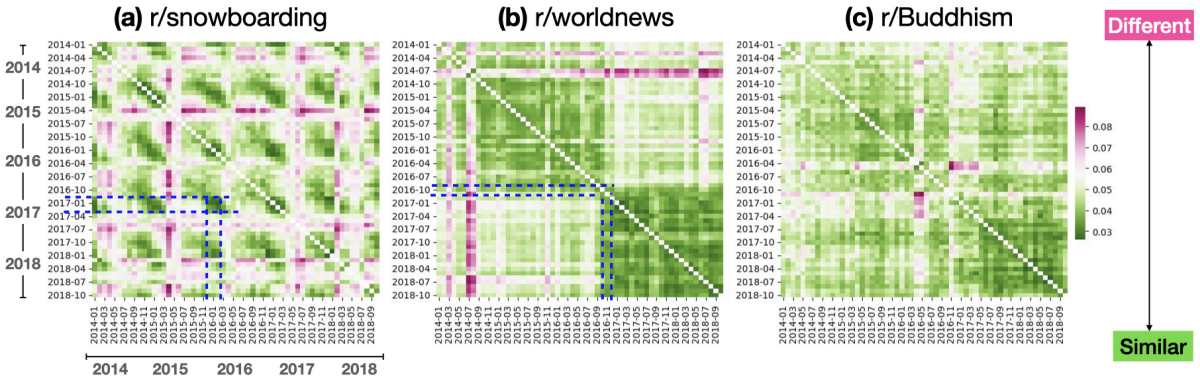

Figure 1: Comment-based EMB distances between months are shown in a heatmap where green represents more similar (small distance) comment language, and red represents less similar (large distance) comment language. Dashed blue bounding boxes are provided for seasonal subreddits *r/snowboarding* and *r/worldnews* to exemplify highly similar pairs of months.

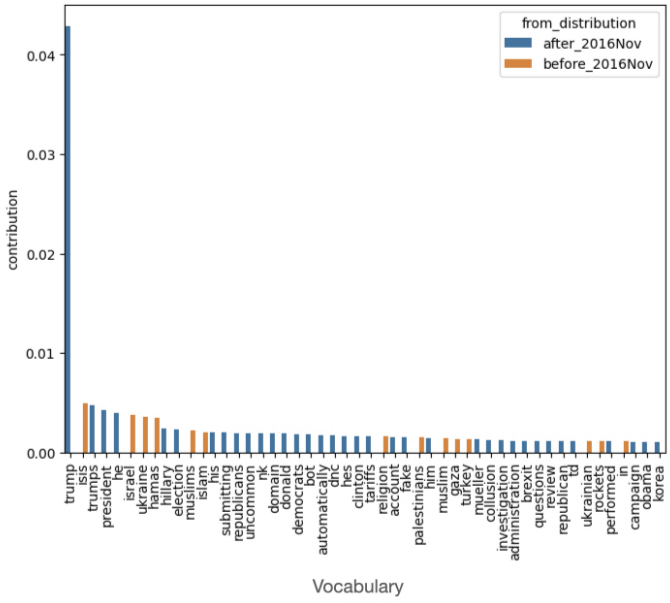

Figure 2: Trump is the unigram most contributing to *r/worldnews* comment seasonality.

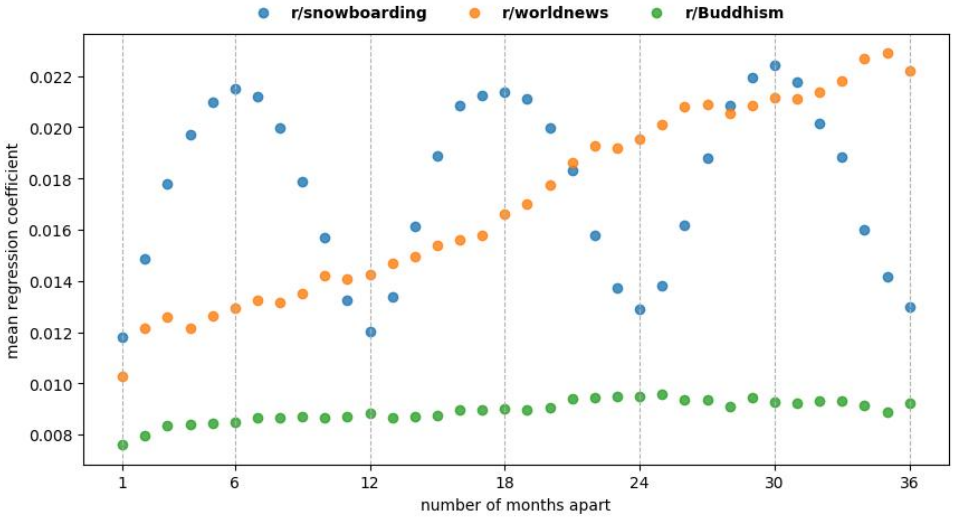

Figure 3: Regression-based seasonality visualization across subreddit case studies.

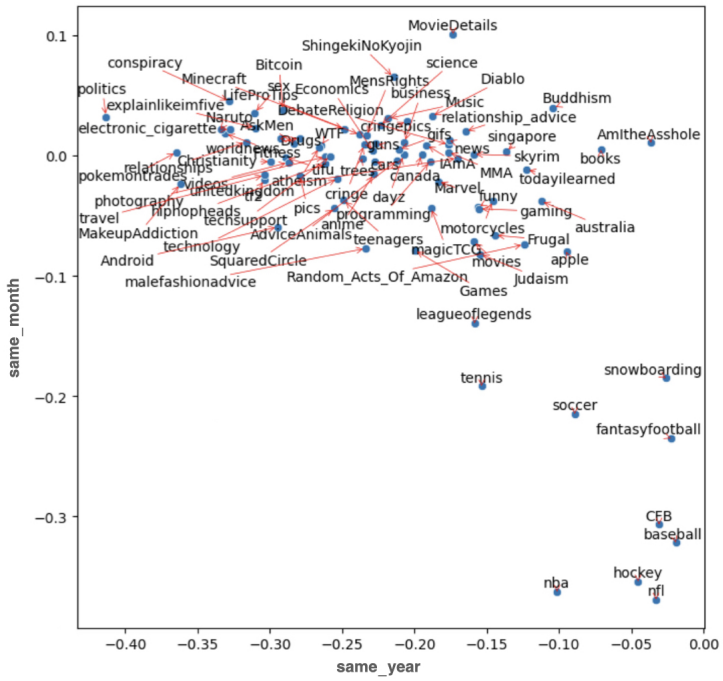

Figure 4: Regression-based seasonality visualization across 84 subreddits. Sports subreddits, exhibiting strong seasonal patterns, are the outliers.