CRII: SaTC: Moderating Effects of Automation on Information Transmission in Social Forums

PI: Jake Ryland Williams, Drexel University Informatics

Award Number: 1850014

To support online discourse moderation and human comprehension of online information, we aim to develop diagnostic applications that allow users and operators to navigate the information environment for themselves, utilizing machine learning (ML)-based software and minimalist, publicly-available data and analytic features.

Key Areas of Focus

- Social bot identification and support labeling as features for news veracity evaluation
- Development of a cross-platform framework for analysis of conversational content
- Controllability of semantic bias in NLP technologies detection and mitigation
- System-level evaluations of social bot impact on information integrity
- Detection of coordinated campaigns and bot networks
- Design through computationally-inexpensive machine learning approaches
- Presenting information to support users in evaluating information from online sources

Developing systems to process diverse social content

Posting Behavior Across Social Media Platforms

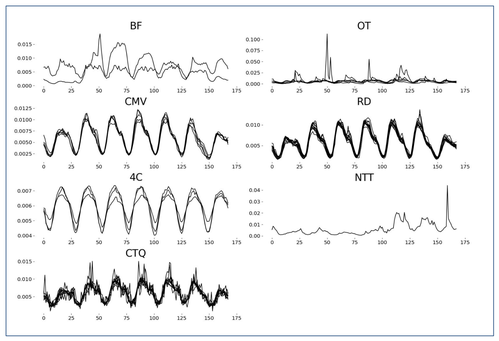

Figure 5: Distribution of the posting time behavior within a week, stratified by dataset. Within a single small multiple, each line corresponds to a different year's data. On the x-axis, 0 is Sunday at midnight. [8]

Data. A principle goal with this CRII project has been to develop a system capable of cross-platform analysis. For this, we have built a common data model for a diversity of discourse environments, which all present variations on conversational structure. To manage these differences, we've produced the *pyconversations* module for Python [9], which provides a common software framework for interacting with conversational data. To build this common data model and interaction framework—as well as the project's other developments now being built on top—our work includes developing and processing for the following data sets (posting behavior is shown in **Figure 5**):

From **Twitter**, threads branching from NewsTweet's embedded tweets are called NewsTweetThreads (**NTT**). A number of user timelines identified from suspicious following activities **[2–3]** were harvested for their quote tweets, and these are referred to as the Coordinated Targeting Quotes (**CTQ**).

From **Facebook**, The BuzzFace dataset [0] records public discourse present on Facebook posts that were fact-checked by BuzzFeed (**BF**); and was augmented with an auxiliary collection of political/news-oriented Facebook page discourse, referred to as Outlets (**OT**).

From **Reddit**, Change My View (**CMV**) is a externally produced dataset that explores the sub-reddit r/ChangeMyView exhibits user posts of opinions alongside challenges made to them to, as the name suggests, change their view. The Reddit Dialog (**RD**) data set consists of Reddit conversations from 3 sub-reddits: r/news, r/worldnews, and r/politics and spans their creation up to January 2019, originally produced by the developers of DialoGPT to train their transformer-based language model.

From **4chan** (**4C**), an ad hoc collection of boards are tracked: news (/news/), history (/his/), science (/sci/), technology (/g/), politically incorrect (/pol/), and paranormal (/x/).

Episodes of Social Media Use

Leveraging *Ancient Accounts* (Accounts created prior to 2009) on Twitter **[2–3]**, information regarding patterns of human social media use can be analyzed to understand an initial research question: ***are there typical patterns of social media use across time***. Grouping tweets within a user timeline by timestamps, distinct episodes of social media use can be modeled for a given user. These episodes are user specific, motivated by the user's own patterns/frequency of social media use. Preliminary investigations identify spikes in social media usage, including *wake up events* where long dormant accounts begin tweeting again, at times correlating with election cycles and the Covid-19 pandemic. In addition to organic *wake up events* where users appear to be seeking to participate in the social conversation, *suspect wake up events* can be seen (**Figure 6**) where users suddenly begin exhibiting an extraordinary increase in social media activity.

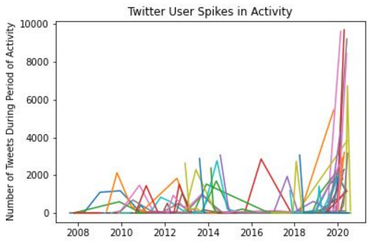

Figure 6: Tweet counts for 51 *ancient accounts* that demonstrated frequent periods of inactivity followed by an extreme spike in activity

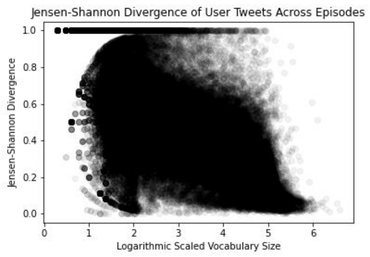

Figure 7: Evaluation of the Jensen-Shannon Divergence for the vocabularies of tweets across social media episodes of use for 116,522 *Ancient Accounts*

When a new episode of social media use occurs, is the same human in control of the account?

Multiple methods are being used to evaluate this question including: A) evaluation of tweet frequency and temporal statistics within an episode of social media use and B) vocabulary similarity and divergence across episodes within a user's social media use. **Figure 7** shows an evaluation of the Jensen-Shannon divergence across adjacent episodes for a subset of the *ancient account* population. This figure depicts two interesting phenomenon: 1. For a subset of users with an extremely large vocabulary size, the terms used across the episodes are completely distinct; 2. There exist distinct smoothing phenomena for a subset of users with both an extremely large vocabularies and extremely small vocabularies.

This research indicates that the classification of 'bot like behavior' can appear within a users timeline, indicating a *regime change* has occurred from a *human-user* to a possible *bot-user* now controlling the timeline.

NewsTweet: A publicly-accessible data stream integrating social and mass media, whose connected subgraphs form hierarchies of coherent, topical stories.

Social media content *embedding* within online news is a now common practice. These *embeddings* serve as both a means of surfacing the 'voice of the general population' and as the newsworthy item of discussion within an article. Leveraging Google RSS feeds as a generalized news aggregator to feed into the system (**Figure 1**), news is categorized into 8 domains: Business, Entertainment, Health, Nation, Sports, Technology, World, and Headlines. The HTML content for articles for each domain supplied by the Google RSS Feed is processed to find and extract embedded tweets. From the limited embedded tweet metadata, a Twitter API call is made using the tweetID. A secondary datastore is then maintained by leveraging the resulting tweet object's userID as the pipeline can access the users most recent 3,200 tweets. This datastore consists of users whose tweets have been *embedded* in the past and are now continuous tracked for new tweets. This allows for the identification of *embedding* patterns and *newsworthy* users. Evaluation of the users shows that a subset of users (mostly celebrities and well-known organizations) are the most frequently *re-embeded* users, but users with with the most *embeds* for the least tweets receive more *unique embeds*. This indicates that although some users have a high *embed* frequency, this is not necessarily correlating with being the unique topic of an article. **[4]**

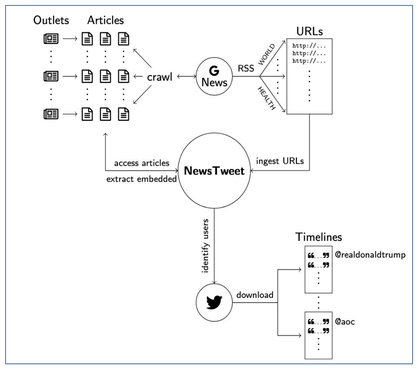

Figure 1:NewsTweet data collection pipeline which 'acquires data on the interaction of embedded tweet content' [4]

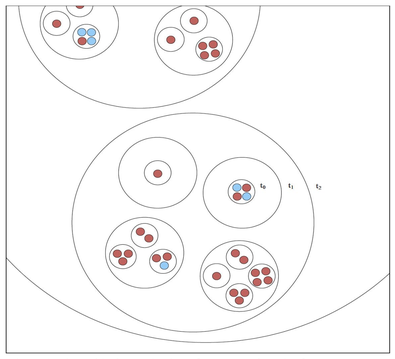

Figure 2: Hierarchical organization of aggregated stories from NewsTweet collection. Red nodes represent articles and blue nodes tweets. Larger circles represent higher levels of aggregation. [13]

The second phase of research focuses on the development of a network construction approach leveraging embedded tweets and identifying connected articles (**Figure 2**). In addition to identifying connected articles, these articles can be semantically segmented into distinct subgraphs to identify topically distinct groups of articles.

Semantic featurization and unsupervised clustering are built on the network structures identified through the connected articles, creating potential for a news aggregator (**Figure 3**) that constructs hierarchies of stories in the news across articles that can help news consumers interactively navigate contexts and hierarchies within emerging news stories. **[13]**

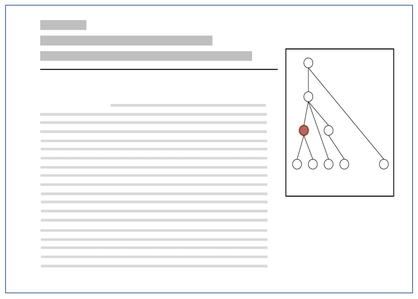

Figure 3: A mock application concept demonstrating story hierarchy navigation interface. Leveraging data acquired through the NewsTweet Pipeline (Figure 1) and aggregated via the approach demonstrated in Figure 2, Users would view and interact with the aggregation hierarchy associated with stories for context. [13]

Broader Implications And Future Work

A Word2Vec's Softmax Factorization

Theorem: Under the log-softmax objective:

$$\mathcal{L}_{soft} = -\sum_{t\in V}\sum_{s\in V} F^m_{t,s}\log\varphi(\vec{u}_t\vec{v}_s), \qquad (9)$$

the Word2Vec algorithm implicitly converges towards a matrix factorization for all non-zero co-occurrences of the form:

$$\vec{u}_t\vec{v}_s^{\,T} = \log\frac{F^m_{t,s}}{f^m_t}, \qquad (10)$$

which is equal to the log-conditional probability matrix of the co-occurrence model.

Bias control, prior to model learning

Bias manifestation in NLP technology is a challenging problem to tackle, and can be dangerously abused in pre-existing. [6] While much research aims to remove bias from NLP systems, we've pioneered methods for probing *data* for bias, capacitating the filtration of biased training data—prior to model learning. **[12]** To demonstrate this, a corpus ***W*** was created from a collection of all pages leading to and from any UK city's Wikipedia page, resulting in a corpus of approximately 200,000 Wikipedia page, and compared to the Google Books corpus' most recent word frequencies (***G***). Given pairs of words in analogies, we've define the *analogical dissonance* as the absolute difference of log-frequency ratios (Δ). Δ is measured over [0,1], and so can be compared across corpora for a **computationally-inexpensive** means of comparing the level of bias between corpora.

Figure 4 shows that ***W***—one-thousand times smaller than ***G***—exhibits less dissonance in (more bias towards) only one single category, namely the subject—cities of England—that was intentionally oversampled in the (biased) construction of ***W***.

Reducing the Computational Complexity and Precision of NLP

NLP technologies are computationally expensive to train, having large numbers of ad hoc model parameters that require iterative updates from (potentially repeated) observation and prediction over data. Not only is this slow and costly, but models learned over large volumes of data can embody surprising, and potentially dangerous biases **[6]**. Operating under the theory that these algorithms simply converge towards maximum-likelihood solutions, we're presently developing mathematical methodologies to *solve* for best-possible NLP models—on a CPU (not a GPU).

Our first approach to this developed the near-bias-free, independent-frequencies model (IFM) **[12]**—which assumes word occurrences are independent—and its most-naive implementation, as the *EigenNoise* representation. **[11]** This work progressed to our discovery of a full parameterization to language-model-based representation via the Word2Vec algorithm's softmax factorization, presented at left and proven in **[12]**.

As an unexpected outcome of defining the IFM, we uncovered an explanation for the well-known 'linear semantics' of analogies, whose mathematical basis we refer to as the frequency-ratios property. We've set this property to define a measure that directly quantifies the extent to which a data set is semantically representative of a collection of analogies, with respect to the word vectors that a representation learning algorithm would learn, as discussed below at left.

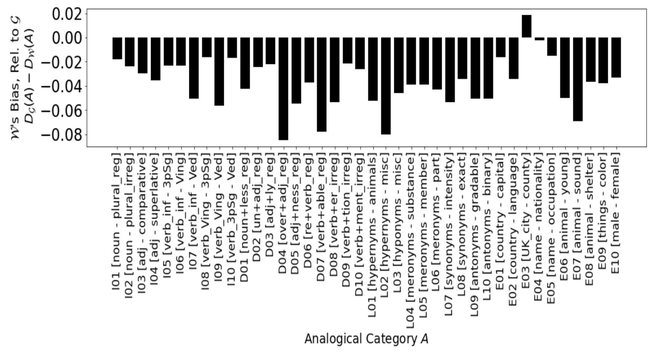

Figure 4: Comparison of the dissonance towards the different BATS analogy categories for the Google Book corpus, ***G***, and a much smaller corpus of Wikipedia articles that connect to pages discussing the UK cities. Positive bars indicate categories towards which ***W*** is more biased, i.e., which contain analogies that ***W*** supports more. [12]

Directly-Related Works Disseminated

[13] Modeling Emerging News Stories Across Digital Publications and Social Media. Dissertation by M. I. Mujib (Advised by J. R. Williams), Ph.D. in Information Science (2022).
[12] To Know by the Company Words Keep and What Else Lies in the Vicinity. H. S. Heidenreich and Jake R. Williams. Arxiv Preprint (2022).
[11] EigenNoise: A Contrastive Prior to Warm-Start Representations. H. S. Heidenreich and Jake R. Williams. Arxiv Preprint (2022).
[10] Unmixing Documents: The Mixing Law and Resonance in Language. Thesis by D. Solano-Oropeza (Advised by J. R. Williams), Bachelor's in Physics (2021).
[9] pyconversations: A package for representing conversations as DAGs for visualization, analysis, and pre-processing. Python Package Index (2021).
[8] Look, Don't Tweet: Representation Learning and Social Media. Thesis by H. R. Heidenreich (Advised by J. R. Williams), Bachelor's in Computer Science (2021).
[7] The Mixing Law and Experiments in Document Malformation J. R. Williams and D. Solano-Oropeza. Fourth Northeast Regional Conference on Complex Systems (2021).
[6] The Earth Is Flat and the Sun Is Not a Star: The Susceptibility of GPT-2 to Universal Adversarial Triggers H. S. Heidenreich and J. R. Williams. AAAI/ACM Conference on AI, Ethics and Society (2021).
[5] A general solution to the preferential selection model. J. R. Williams, D. Solano-Propeza, and J. R. Hunsberger. Arxiv Preprint (2020).
[4] NewsTweet: A Dataset of Social Media Embedding in Online Journalism. M. I. Mujib, H. S. Heidenreich, C. J. Murphy, G. C. Santia, A. Zelenkauskaite, and J. R. Williams. Arxiv Preprint (2020).
[3] Investigating Coordinated 'Social' Targeting of High-Profile Twitter Accounts. H. S. Heidenreich, M. I. Mujib, and J. R. Williams. Arxiv Preprint (2020).
[2] Investigating Coordinated 'Social' Targeting of High-Profile Twitter Accounts. H. S. Heidenreich, M. I. Mujib, and J. R. Williams. International Conference on Computational Social Science (2020).
[1] Detecting Social Bots on Facebook in an Information Veracity Context. G. C. Santia, M. I. Mujib, and J. R. Williams. International AAAI Conference on Web and Social Media (2019).
[0] BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos. G. C. Santia and J. R. Williams. International AAAI Conference on Web and Social Media (2018).