Russian Trolls & Emoxicon

Tara L. Valladares1 & Hudson Golino1

1 University of Virginia

Author note

Correspondence concerning this article should be addressed to Tara L. Valladares, Postal address. E-mail: [tls8vx@virginia.edu](mailto:tls8vx@virginia.edu)

Abstract

abstract

Russian Trolls & Emoxicon

# Introduction

Text comprises the majority of information in the world. By and large, human communication relies heavily upon digital messages sent through email, text messaging, and social media platforms. Developing methods that can analyze unstructured text is therefore critical in numerous fields, including psychology, political science, and economics. Specifically, the need for methods of text analysis grows as the broad spectrum influence of social media platforms such as Facebook and Twitter increases. As social media platforms are frequent sites of both information and disinformation campaigns, efforts to identify and understand these processes are limited by the available tools. In this paper, we will introduce a new method of emotion extraction and describe an application to a dataset.

## State Sponsored Disinformation and Russian Trolls

State sponsored disinformation campaigns are not a new phenomenon. Purposeful efforts to spread false information and "fake news" have been documented since ancient Egypt and Rome. With the rise of the internet and social media, such campaigns can be conducted at a level and size that was previously impossible. The Russian state commonly employs disinformation campaigns in an effort to influence foreign politics. These campaigns utilize a variety of strategies to sow discord, including online media manipulation, cyber-espionage, and falsely inciting protests in the real world (Linvill et al., 2019; Richey, 2018).

The term *trolls* or *sock puppet* accounts is commonly used to described social media accounts that claim to be some person from a certain background or position, while actually being under direct control of a hidden third party. This is similar but in contrast to *bots* which perform similar actions to trolls but are directed by a computer program rather than a real person. Both trolls and bots have been extensively used by Russia, and other governments, to manipulate social media and political discourse around the world.

During the 2016 U.S. presidential election, Russian troll farms posed as U.S. natives on Twitter in an attempt to influence political discourse and opinion. This eventually lead to the United States Justice Department inditing 13 Russian individuals and three Russian entities on charges relating to interference in the presidential election (Barrett, Horwitz, & Heiderman, 2018). The Internet Research Agency (IRA), a company based in St. Petersburg, was among the companies charged and has been a primary focus of investigations into state-sponsored Russian interference in U.S. politics (Permanent Select Committee on Intelligence, 2018). After the Justice Department released the Twitter handles associated with the IRA, Clemson university researchers Darren Linvill and Patrick Warren downloaded all tweets associated with these accounts and made the dataset publicly available for researchers (Linvill & Warren, 2019). The full dataset consists of nearly 3 million tweets from 2,800 accounts.

The content of the Russian troll tweets ranges widely. (Linvill & Warren, 2019) documented four major categories of English-language accounts that were active before the 2016 election: right trolls, left trolls, news feeds, and hashtag gamers. The majority of tweets from all accounts were deemed innocuous camouflage tweets; that is, the tweets were not political, but were instead contained mundane interactions with followers or references to popular culture (Linvill et al., 2019). Political tweets from right-wing accounts largely supported Donald Trump as the presidential nominee and attacked Hilary Clinton. Political tweets from left-wing accounts largely attacked Trump, while less frequently either attacked or supported Clinton. There were apparent differences in how the account categories tweeted, though. For example, the right-trolls had less camouflage tweets than the left-trolls. It was estimated that over half of tweets from left-troll accounts were camouflage tweets, compared to less than 15% of tweets from right-troll accounts (Linvill et al., 2019). This suggests that left and right troll accounts were employing different strategies in their attempts at political influence.

Because both typical human communication and disinformation campaigns commonly occur on mediums characterized by brief segments of texts (e.g., tweets), there is a great need for flexible, low-computation methods of emotion detection that reliably work on relatively short documents.

## Brief history of emotion extraction

An enormous amount of work has been dedicated to extracting sentiment and emotion from text. As is common in many developing fields, the nomenclature of techniques can be unclear. Sentiment analysis is often used as a catchall term to describe the measurement of any feeling in text, no matter how complex. Here, we will separate the extraction of emotion from that of sentiment. Sentiment analysis is the process by which sentiment, i.e., positive and negative opinions, is extracted from text. Emotion detection focuses on the extraction of emotional categories, such as happiness, anger, or pride. The distinction between the sentiment and emotion can lie on a continuum, depending how features are extracted and utilized. Competing theories and methods may conceptualize emotions as distinct categories or as lying on related continuums. When categorical, the emotion categories generally include anger, disgust, fear, joy, sadness, and surprise, and sometimes trust and anticipation (Ekman, 1992; Hirat & Mittal, 2015; Mohammad, 2016; Strapparava & Mihalcea, 2008). When continuous, popular theories describe emotions as lying on three dimensional vectors constructed from arousal, valence, and dominance (Russell, 2003).

### Emotion Lexicons.

Lexicons form the basis of most methods of emotion detection. Lexicons are dictionaries that contain features (such as words or punctuation), the category they belong to, and their designated score. Lexicons vary in their size, complexity, and focus. They may be manually created through annotation or created automatically through machine learning.

Lexicon-based analyses are limited by the lexicon’s size and complexity. In general, the larger the lexicon, the more robust the scoring. Smaller lexicons may be poor matches to a target text if there is little overlap in words. Significant bias can occur when a target text is poorly matched to the lexicon. For this reason, domain-specific lexicons can give better results than general lexicons, though the creation of lexicons for every domain and genre is a monumental task. The majority of lexicons are created or seeded through hand annotation (Mohammad, 2016). While human generated lexicons have validity advantages, they will be naturally limited in size and by responder biases.

### Bag-of-words.

Bag-of-words (BOW) is a method of cleaning and organizing text that is commonly used alongside lexicons. In BOW, all grammar and word placement is discarded to create an orderless “bag” of word counts. BOW is thus a method of extracting the word frequency. Lexicon-scored and frequency data can be analyzed through simple sum scores or with more complex methods. While BOW and lexicons can serve in standalone analyses, machine learning frequently makes use of BOW and lexicons to create features.

Though relatively crude, bag-of-words can produce compelling results on its own or in combination with machine learning techniques. (Da Silva, Hruschka, & Hruschka Jr, 2014) found that BOW combined with standard machine learning algorithms produced accuracy rates of tweet sentiment classification around 70-80%. There, the addition of sentiment lexicons always improved classification rates.

## Aims

The purpose of the current study is to introduce a novel method of emotion detection, termed Emoxicon, and detail its application to a dataset of tweets sent by Russian Trolls.

$$\[expand$$

]

# Methods

## Emoxicon

Emoxicon is a method of analyzing the emotional content of text. The Emoxicon method primarily consists of two parts: emotion scoring and Rasch modeling.

### Lexicon Scoring.

Lexicon-based analyses are limited by the lexicon’s size and complexity. In general, the larger the lexicon, the more robust the scoring. Smaller lexicons may be poor matches to a target text if there is little overlap in words and significant bias can occur when a target text is poorly matched to the lexicon. We used the DepecheMood++ lexicon created by (Araque, Gatti, Staiano, & Guerini, 2018). DepecheMood++ is an emotion lexicon that contains roughly 37 thousand terms and their associated probability weights for 8 emotion categories. The lexicon was created by scraping data from the news website Rappler which features a native "*Mood Meter*" widget on each article. Readers of the site are encourage to select one of eight reactions (Afraid, Amused, Angry, Annoyed, Don’t Care, Happy, Inspired, or Sad) to the article. DepecheMood++ has several features that make it particularly useful. First, DepecheMood++ was naively crowdsourced; it did not rely on hand-annotation by knowledgeable participants. Second, DepecheMood++ is much larger than other commonly used lexicons. Previous work has shown that the Depechemood++ lexicon performs as well or better than other emotion lexicons (Araque et al., 2018; Staiano & Guerini, 2014)

A simple bag-of-words method is applied to score the emotional content of a given document. In the original lexicon, each word is assigned a weight for each of the eight dimensions. Here, the lexicon is condensed; each word in the lexicon becomes associated with the emotion category with the highest probability weight. Then, each tweet is assigned a sum score of how many words are most highly associated with each emotion category.

### The Rasch Model.

The Rasch model is used to evaluate the relative probability of the an emotion weight appearing in a given tweet. By using the Rasch model, we are able to analyze the fit of the emotion weights to the document. The Rasch model originated in the field of psychometrics to relate scales and questionnaires to the underlying latent trait (Rasch, 1960). The Rasch model is a logistic model that places persons and items on the same scale to calculate the probability of a given person endorsing a given item. Fundamentally, the Rasch model can be used to generate latent trait scores for entities that produces a set of responses determined by an underlying latent trait. The basic form of the Rasch model for dichotomous items, as adapted from (Rasch, 1960), is:

Where represents the trait level of person and represents the difficulty or location of item . In Emoxicon, tweets take the place of persons and emotional scores are represented as items. Each tweet is scored based on how many words are present from each emotion category. A mean split was performed on the emotion word counts within each emotion category to dummy code each tweet as 1 (high) or 0 (low) for each category. For example, a tweet with a relatively high amount of "Happy" words would receive a 1 for the "Happy" item. The dichotomized scores are run through the Rasch model to produce trait scores for the tweets and ‘item’ locations for the emotion categories.

## Russian Trolls Dataset

The Russian Trolls Twitter Dataset was collected by XXX.

From the total dataset, we included only English-language trolls labeled as Left-wing ("Left") or Right-wing ("Right"). Retweets, manual reposts (i.e., tweets that begin with "rt"), and duplicate tweets were removed.

paragraph about why we focused on authored tweets

Our final sample consisted of 420447 tweets from 622 right-wing accounts ( = 347489) and 227 left-wing accounts ( = 72958). Among twitter handles with at least 30 tweets, there were 429 right-wing accounts ( = 345489) and 163 left-wing accounts ( = 72162).

Words that were hypothesized to specifically associate with the topic and not than emotional content were removed. For example, the words "Hillary," "Clinton," and "Trump" are present in the Depechmood++ lexicon. These words were not scored as we believed that such words did not accurately represent their respective emotion categories in this circumstance. That is, Right and Left trolls may use the word "Trump" just as frequently, but that word does not have the same relationship to the tweets true true emotional content for each group. As these words were used frequently, they also would artificially inflate their respective category "easiness". See Appendix 1 for a full list of removed words.

Distribution of tweets (table?)

Tweets dated from XXXX to XXXX.

To analyze this dataset, we fit one overall Rasch model containing all tweets from all accounts, as well as individual models for each twitter account.We examined consistency of the category difficulties by fitting individual Rasch models for each twitter handle, resulting in 592 separate Rasch models. As mentioned, only accounts with at least 30 tweets were included to sufficiently fit the individual Rasch models (Linacre, 1994). Because the mean of either the person or item parameters must be set to zero before estimation, item parameters are not directly comparable across separate Rasch models. Therefore to compare the relative difficulties between models, item parameters from the were transformed into ranks from 1 (easiest) to 8 (hardest) within each model. The distribution of the ranks could then be examined for consistency within emotion categories and between Right and Left trolls. If an individual account did not use any words from a specific emotion category, that emotion category was given the rank of 8 (hardest to endorse). This occurred most frequently for "Happy" (n = xxx) and then xxx.

# Results

## Overall Rasch Model

# library(emoxicon)

We fit a Rasch model

model fit person fit item fit item distribution person distribution

## Individual Rasch Models

We fit 592 Rasch models for each of the individual twitter handles with more than 30 authored tweets.

person/item distribution

model fit

person fit s

item fit

### Category Ordering.

For all individual models, the hardest category was XXX and the easiest category was XXX (Table XX).

### Gini Index (Emodiveristy).

### Mutual Information.

### Prediction.

# Discussion

## Future Directions

## Limitations

## Conclusions

# Words removed from Lexicon

hilaryhillaryrussiarussiantrumpbernieclintonrt

Araque, O., Gatti, L., Staiano, J., & Guerini, M. (2018). DepecheMood++: A bilingual emotion lexicon built through simple yet powerful techniques. Retrieved from [https://arxiv.org/abs/1810.03660](%20https://arxiv.org/abs/1810.03660)

Barrett, D., Horwitz, S., & Heiderman, R. (2018). Russian troll farm, 13 suspects indicted in 2016 election interference. *Washington Post*.

Da Silva, N. F., Hruschka, E. R., & Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, 170–179.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169–200.

Hirat, R., & Mittal, N. (2015). A survey on emotion detection techniques using text in blogposts. *International Bulletin of Mathematical Research*, *2*(1), 180–187.

Linacre, J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 328.

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). “THE RUSSIANS ARE HACKING MY BRAIN!” Investigating russia’s internet research agency twitter tactics during the 2016 united states presidential campaign. *Computers in Human Behavior*.

Linvill, D. L., & Warren, P. L. (2019). Troll factories: Manufacturing specialized disinformation on twitter. Retrieved from [http://pwarren.people.clemson.edu/Troll\_ Factories\_v2\_Linvill\_Warren.pdf](http://pwarren.people.clemson.edu/Troll_%20Factories_v2_Linvill_Warren.pdf)

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 201–237). Elsevier.

Permanent Select Committee on Intelligence. (2018). Schiff statement on release of twitter ads, accounts and data. Retrieved from <https://democrats-intelligence.house.gov/news/documentsingle.aspx?DocumentID=396>

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

Richey, M. (2018). Contemporary Russian revisionism: Understanding the Kremlin’s hybrid warfare and the strategic and tactical deployment of disinformation. *Asia Europe Journal*, *16*(1), 101–113.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145.

Staiano, J., & Guerini, M. (2014). Depechemood: A lexicon for emotion analysis from crowd-annotated news. *arXiv Preprint arXiv:1405.1605*.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 acm symposium on applied computing* (pp. 1556–1560).