Using Deep Networks and Transfer Learning To Address Disinformation

Numa Dhamani, Paul Azunre, Jeffrey L. Gleason , Craig Corcoran,
Garrett Honke, Steve Kramer, Jonathon Morgan

BACKGROUND

Can semantic classification of natural language be used as a tool for the detection of inflammatory, inauthentic, or otherwise nefarious communication?

An ensemble pipeline composed of a character-level convolutional neural network (CNN) and a long short-term memory (LSTM) network is tested as a general tool for addressing a range of disinformation problems.

We demonstrate that transfer learning from one domain to related (supervised and unsupervised) tasks is highly effective and character-level input representation is critical—we attribute this effectiveness to the properties of the architecture that allow processing of messy nature of social media communication in the face of a lack of labeled data and multi-channel tactics of influence campaigns.

ARCHITECTURE

We focused on an ensemble of two coupled deep neural networks[1]—a network that encodes each individual sentence and a network that classifies the entire document.

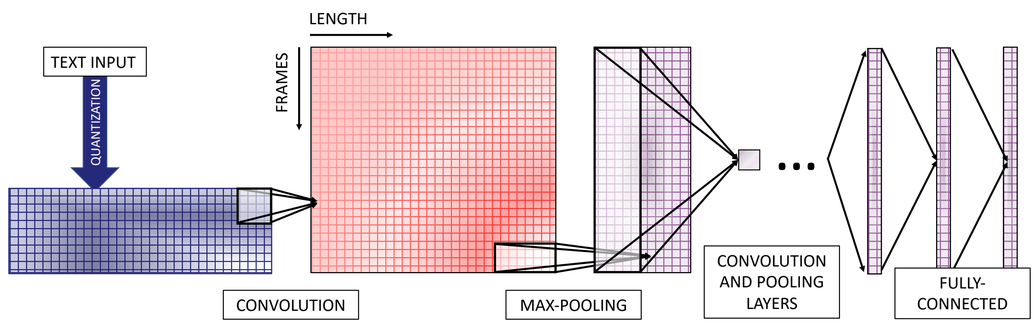

Figure 1. Diagram of character-level convolutional neural network model architecture. Figure adapted from Zhang et al., 2015.

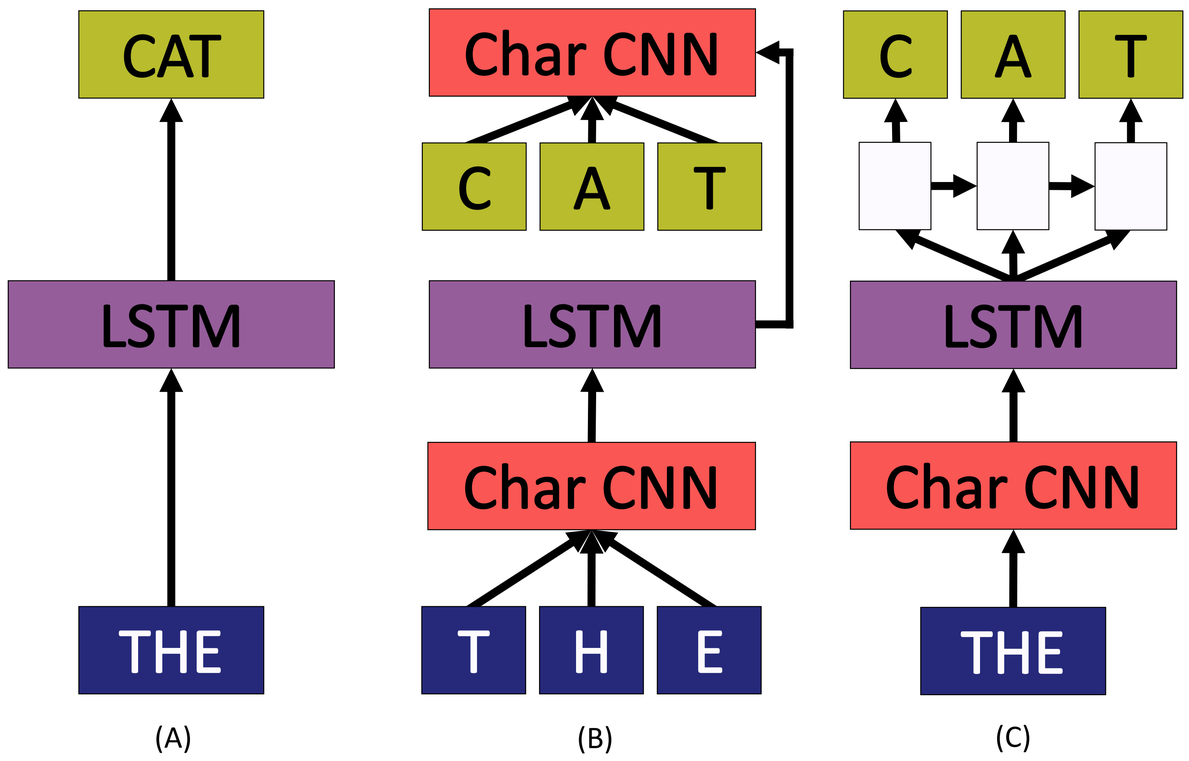

Figure 2. Candidate word and character-level ensemble models for text generation and classification. This work focused on the architecture in Panel B. Figure adapted from Józefowicz et al., 2016.

[1]Source code: https://github.com/NewKnowledge/simon

RESULTS

Spam Emails—peer-to-peer social engineering attacks

Training

Binary: friend, foe

Data: Enron email dataset, 419 spam fraud corpus, email abuse dataset (acquired from NASA JPL)

Training test binary accuracy: **96.14%**.

Transfer Learning

Multi-class: friend, 419 scam, malware, credential phishing, phishing training, propaganda, social engineering, spam

Test accuracy: **93.25%**.

Review Bombing—campaigns to threaten the integrity and good reputation of a product, company, or a brand.

Initial Results

Binary: on-topic, off-topic

Data: movie reviews collected from crowdsourced review website

Test accuracy: **99.5%**.

Political Sentiment—dissemination of false narratives with broad implications for society by piggybacking on polarizing content or current events.

Results

Multi-class: pro, neutral, anti

Data: social media posts discussing former Starbucks CEO Howard Schultz' potential candidacy for the 2020 U.S. presidential election.

Test accuracy: **88.10%**.

Conversation Clustering—examining the structure and trends of online discourse provides critical context for detecting and understanding large-scale disinformation campaigns *(Figure 3).*

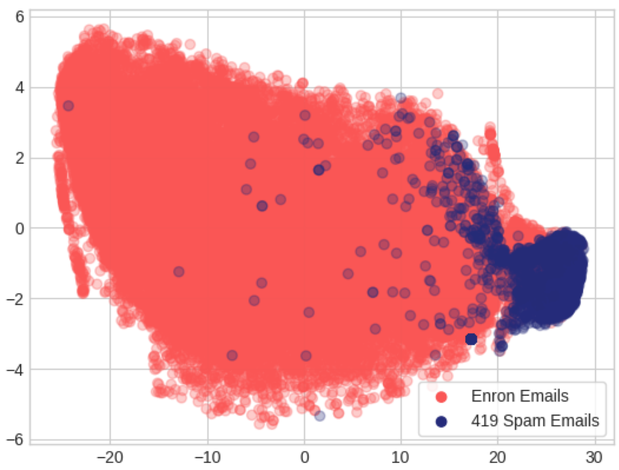

Figure 3. t-SNE embedding of model features from the Enron and 419 spam email datasets.

FUTURE DIRECTIONS

Semantic text classification is a critical technique for identifying nefarious communication; yet, it is only one component of a broader toolkit needed to effectively combat disinformation.

The representation produced by this ensemble network can be used for tracking language usage across communities and changes in language over time, for determining the early or "originator" communities that show distinctive language patterns.

Ongoing work examines these dynamics of language usage across communities and social platforms.