Multimodal Prediction of Suspicious News Types on Twitter

Social media platforms are a convenient, fast, and popular source of information for keeping up-to-date on current events. About 62% of American adults report using social media like Twitter and Facebook as a means of staying informed. Not every news report, however, is vetted and verified before posted. Untrustworthy accounts exploit and deceive these users in order to achieve a hidden purpose (e.g. opinion manipulation or web traffic attraction) and reply on the underlying connected network of users as a way to spread their stories. Separating suspicious messages from reliable requires much effort, in addition to being time consuming and tiresome. Therefore, the goal of this work is to construct a joint image and text model in order to classify social media posts as mainstream or alternative news, as well as predict three types of alternative news: disinformation, clickbait, or propaganda. In addition to this task, our research aims to answer these questions: 1) Can we improve the accuracy of existing methods by including images which have previously been ignored? 2) Can you measure what signals in social media posts determine the class of a post e.g., text, linguistic cues or images?

Our model consists of two subnetworks trained jointly for text and images. The two subnetworks are merged using the late fusion technique. The text side contains an embedding layer initialized using GLoVe embeddings followed by a Long Short-Term Memory (LSTM) layer. We extract linguistic cues on a tweet level to enrich the inputs to the LSTM layer. Linguistic features encode specifically for language that contains biased terms, moral foundation cues, and the subjectivity of a tweet.

Images are incorporated via a standard feed-forward network after transforming into 2048-dimensional feature vectors obtained using ResNet architecture trained on ImageNet. For a completely text based approach, we implemented Tensorflow's Object Detection model to identify objects within the supplied images. From these objects, we applied scikit-learn's Tfidf-transformer to represent each image as a feature vector. We evaluated our model using different feature combinations and different architectures. We experimented with 2048-dimensional image encodings, tfidf representations of image object features, and text embeddings combined with linguistic cues.

We designed two evaluation experiments: separating verified and deceptive news and distinguishing between the three types of deceptive news. Our experiments are ongoing and preliminary results on our binary task show an 83% accuracy achievable with a full textual model, 94% accuracy with the addition of linguistic cues, and 91% accuracy with text, images, and linguistic features. When predicting exclusively with images, we achieve 73% accuracy using the object tags from the Tensorflow model and 82% accuracy with the 2048-dimensional feature vectors. We are still conducting and gathering results for the multiclass task.

Recent work in multimodal language modeling has improved how images and text can be represented in the same space, giving increased powerful meaning to embedded representations and improving performance on tasks such as image captioning, machine translation, and natural language inference.[1][2] Previous studies in deceptive news detection have relied solely on textual representations for classification.[3],[4] Our unique approach utilizes state of the art multimodal methods and applies them to the difficult task of detecting falsified news stories.

1. Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models." *arXiv preprint arXiv:1411.2539* (2014).
2. Vendrov, Ivan, et al. "Order-embeddings of images and language." *arXiv preprint arXiv:1511.06361* (2015).
3. Wang, William Yang. "" Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection." *arXiv preprint arXiv:1705.00648* (2017).
4. Volkova, Svitlana, et al. "Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2017.