Explaining Multimodal Deceptive News Prediction Models

Svitlana Volkova[1], **Ellyn Ayton**[1], **Dustin L. Arendt**[2], **Zhuanyi Huang**[2], **and Brian Hutchinson**[1,3]
Data Sciences and Analytics Group[1], Visual Analytics Group[2]
National Security Directorate, Pacific Northwest National Laboratory
902 Battelle Blvd, Richland, WA 99354
Computer Science Department, Western Washington University[3]
516 High Street, Bellingham, WA 98225
{firstname.lastname}@pnnl.gov

Abstract

In this study we present in-depth quantitative and qualitative analyses of the behavior of multimodal deceptive news classification models. We present several neural network architectures trained on thousands of tweets that leverage combinations of text, lexical, and, most importantly, image input signals. The behavior of these models is analyzed across four deceptive news prediction tasks. Our quantitative analysis reveals that text only models outperform those leveraging only the image signals (by 3-13% absolute in F-measure). Neural network models that combine image and text signals with lexical features e.g., biased and subjective language markers perform even better e.g., F-measure is as high as 0.74 for binary classification setup for distinguishing between verified and deceptive content identified as disinformation and propaganda. Our qualitative analysis of model performance, that goes beyond the F-score, performed using a novel interactive tool ErrFilter[1] allows a user to characterize text and image traits of suspicious news content and analyze patterns of errors made by the various models, which in turn will inform the design of future deceptive news prediction models.

Introduction

Social media plays a significant role in modern day information dissemination, owing to its ubiquity and the ease of sharing. Unfortunately, the same qualities that allow information to be rapidly and widely spread also allows social media to be used to deceive. Interest in the manual and automatic detection of suspicious news across social platforms including Facebook, Twitter, YouTube and even WhatsApp has intensified recently (Lazer et.al. 2018).There has been an increased focus on tracking the spread of misinformation through social networks and in particular, during times of crisis (Starbird 2017; Vosoughi, Roy, and Aral 2018).

One approach to suspicious news prediction depends on crowdsourcing to identify, flag, and track potentially suspicious news content that requires expert annotations for verification (Kim et al. 2018).These methods are both labor intensive and subject to human limitations (Kasra, Shen, and O'Brien 2018).

Several automated approaches to deceptive news detection have been developed. Many previous studies have relied solely on textual representations as input to machine learning models, using large bodies of text from online news sites and entire articles (Rashkin et al. 2017; Volkova and Jang 2018). The limited availability of lengthy text segments necessitates the addition of more information from the post, e.g. images (Kiros, Salakhutdinov, and Zemel 2014), social network connections or lexical markers of subjective and biased language, in order to accurately predict suspicious news content (Volkova et al. 2017).

Our contributions are two-fold. First, we present a quantitative analysis of classification performance across neural network architectures, input signals and prediction tasks. Second, we offer a series of qualitative analyses of model performance on these tasks, identifying characteristics of different categories of deceptive news and analyzing cases of agreement and disagreement between model predictions. Taken together, our findings will provide understandings of multimodal content online (both images and text), and as a result inform and inspire future approaches for identifying the credibility of news content shared online.

Data

Our data collection effort supports four predictive tasks that aim to characterize content retweeted from suspicious and verified news accounts on Twitter at different levels of credibility and author's intent. While many definitions exist, we define types of deceptive news content as follows:

- Disinformation contains fabricated information to intentionally mislead the audience.
- Propaganda is defined as the art of influencing, manipulating, controlling, changing, inducing, or securing the acceptance of opinions, attitudes, action, or behavior.
- Hoaxes include scams or deliberately false or misleading stories.
- Conspiracies are an effort to explain some event or practice by reference to the machinations of powerful people, who attempt to conceal their role.
- Clickbait presents content for attracting web traffic.
- Satire includes statements of which the primary purpose is to entertain.

[1] https://github.com/pnnl/errfilter

Table 1: The number of unique retweets with images for each class and prediction task separated by deceptive intent.

	Task 1	Task 2 – 3	Task 4
Intent to Deceive	Low	High	Mixed
Clickbait	620	–	8,147
Conspiracy	622	–	4,833
Hoax	611	–	611
Satire	632	–	2,890
Disinformation	–	18,601	18,601
Propaganda	–	19,040	108,799
Verified	–	19,050	353,048
Total	**2,485**	**56,691**	**496,929**

- VERIFIED includes most trusted content e.g., factual information.

Data Collection Through the Twitter public API, we obtain approximately 4.5M tweets in English retweeted from news accounts in 2016 similar to (Volkova et al. 2017). We then remove retweets where either the included image or the text body is an exact match of another in our collection. We additionally remove retweets without images or where the image is no longer available. Our resulting dataset contains approximately 0.5M unique retweets with image content. Table 1 lists the number of retweets for each class and prediction task separated by intent to deceive.

Data Annotation Twitter news account annotations – propaganda, hoaxes, clickbait, conspiracy, satire as well as disinformation come from earlier work (Volkova et al. 2017; Volkova and Jang 2018). Each retweet in our collection mentions a news source. Thus, we focus on news sources rather than individual false stories as recommended by (Lazer and others 2018) and propagate news account class labels to the retweets we collected.

Limitations We acknowledge the limitations of this approach. For example, not all retweets from the account identified as disinformation may not contain disinformation. Moreover, class boundaries are not clear e.g, propaganda, disinformation and conspiracy are usually confused by the annotators and cause lower annotation agreement. In addition, one account can have multiple class labels e.g., propaganda and conspiracy. However, tweet-level annotations of content credibility is labor intensive and susceptible to human biases (Karduni et al. 2019). Crowdsourcing methods require expert annotation for news content verification, and agreement across annotators is low (Kim et al. 2018).

Methodology

Predictive Task Definitions In our first task, we categorize the differences between the four classes with the lowest intent to deceive – clickbait, conspiracy, hoax, and satire (2,485 posts). The second task distinguishes verified news content from the two most deceptive classes – disinformation and propaganda (56,691 posts). Our third task focuses on a binary classification between verified and suspicious content with the highest intent to deceive – the combination of disinformation and propaganda posts (56,691 posts). Our fourth task is designed to differentiate across all seven types of news (496,929 posts).

Text Signals We apply standard text pre-processing techniques and remove hashtags, mentions, punctuation, numerical values, and URLs from the text content of posts and convert them to text sequences. We then initialize the embedding layer weights with 200-dimensional GloVe embedding vectors pre-trained on 2 billion tweets (Pennington, Socher, and Manning 2014). We filter our vocabulary to contain only the most frequent $10,000$ tokens and initialize out-of-vocabulary (OOVs) tokens with zero vectors.

Lexical Markers To capture *biased* e.g., quotations, markers of certainty, inclusions, and conjunctions and *persuasive* language in tweets e.g., factual data, rhetorical questions, and imperative commands, we extract psycholinguistic features from tweet content using the Linguistic Inquiry Word Count (LIWC) lexicon (Pennebaker, Francis, and Booth 2001). Similar to earlier approaches (Volkova et al. 2017; Rashkin et al. 2017), we extract *bias cues* from preprocessed tweets that include *hedges* – expressions of tentativeness and possibility, *assertive verbs* – the level of certainty in the complement clause, *factive verbs* – presuppose the truth of their complement clause, *implicative verbs* – imply the truth or untruth of their complement, and *report verbs*; we also extract *subjectivity cues* by relying on external publicly available subjectivity, and positive and negative opinion lexicons. Our lexical cue vectors are $2,664$ dimensional.

Image Signals We extract $2,048$-dimensional feature vectors from images using the 50 layer ResNet architecture (He et al. 2016) pre-trained with ImageNet weights. Specifically, our feature vectors are the last fully connected hidden layer. We use these vector representations as input to the image, and joint text and image models.

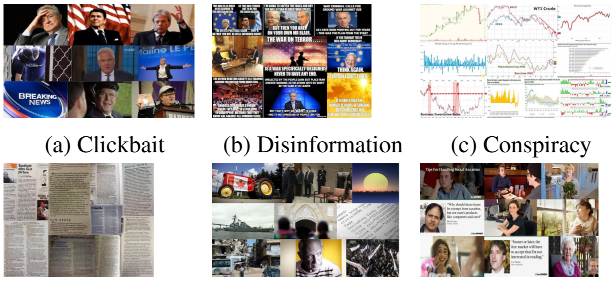
Deceptive News Prediction Models

Relying on our earlier work (Volkova et al. 2017), we experiment with several neural network architectures to evaluate the contribution of different predictive signals for classifying types of suspicious and verified news content.

- IMAGEONLY model relies only on 2,048-dimensional vector representations extracted from image content.
- TEXTONLY model relies on text representations encoded using GloVe embeddings and passed to a Long Short-Term Memory (LSTM) layer.
- TEXT+LEX model incorporates psycholinguistic, bias, and subjective language cues extracted from the tweet.
- TEXT+(LEX+IMAGE) model concatenates vector representations of the image and lexical signals. This combined vector is then passed to a single layer network before we concatenate the last output layer from the TEXTONLY sub-network. We then feed this final concatenated vector to a 2-layer network before making a classification.
- (TEXT+LEX)+IMAGE model mirrors the above model just described except the first concatenation is between the lexical and textual features and the image vector is included in the final concatenation. We include these two similar models in our experiments to analyze the effect of different signal combinations.

Table 2: Classification results across four prediction tasks reported as F1 score obtained using the AdaBoost (AB) and neural network (NN) models. F1 scores for the majority class reported in the parentheses. The highest F1 are highlighted in bold.

Input signals	**Task 1 (0.25)**		**Task 2 (0.33)**		**Task 3 (0.66)**		**Task 4 (0.75)**
	AB	NN	AB	NN	AB	NN	NN
Image	0.457	0.507	0.459	0.462	0.577	0.685	0.137
Text	0.486	0.507	**0.525**	0.596	**0.696**	0.735	**0.267**
Text + Lex	0.478	0.556	0.525	0.590	0.695	0.730	0.216
Text + (Lex + Image)	–	0.562	–	0.591	–	**0.738**	0.220
(Text + Lex) + Image	**0.505**	**0.585**	0.469	**0.598**	0.616	0.737	0.125



(a) Clickbait (b) Disinformation (c) Conspiracy

(d) Hoax (e) Propaganda (f) Satire

Figure 1: Images representative of each deception class chosen among the tweets with the highest confidence scores for the respective classes obtained using ImageOnly model.

Baselines We also consider two baselines against which we will evaluate our models: 1) the majority class and 2) an AdaBoost classifier using decision trees (AB). AB tunable parameters include the type and number of base estimators, and the learning rate. We use Keras with the Tensorflow backend to build, tune, train, and evaluate all models. We tune all of the models on the first prediction task, and then use the chosen configuration for the other tasks. For each experiment we divide data into train 80%, development 10% and test 10%.

Classification Results and Error Analysis

We present the results of the four classification tasks in Table 2. Our experiments reveal improved performance with the use of more than one predictive signal e.g., Text +(Lex+Image) is better than TextOnly and ImageOnly[2] signals.

We observe that the best model (Text +Lex) + Image for predicting suspicious (disinformation and propaganda) vs. verified content yields F1 of 0.738. Models for classifying four types of suspicious news with lower intent to deceive, and disinformation vs. propaganda vs. verified content yield F1 score of 0.585 and 0.598, respectively. Finally, we found that distinguishing across seven types of news sources on a collection of 0.5M tweets is a difficult task – the best model is a textOnly model that yields F1 of 0.267.[3] The following sections detail our qualitative analysis that goes beyond the F score, and focuses on analyzing patterns of errors made by the various models.

Image Characteristics of Model Predictions

We perform several qualitative analyses to investigate where our models fail or succeed to classify tweets in the various classes using a novel interactive tool, ErrFilter[4], that allows us to effectively identify normative and contrastive examples with the goal to inform and improve the accuracy of future deceptive news prediction models.[5] First, we look at image characteristics indicative of each deceptive class. We identify these images from tweets that the ImageOnly model correctly classifies with the highest confidence. Figure 1 shows images from the most confidently predicted tweets. In each class, we can distinguish features uniquely characteristic for that label as shown below.

- *Clickbait* images consist of head shots of politicians or celebrities.
- Graphs and charts are highly indicative of *conspiracy*.
- *Hoax* images contain pictures of newspapers or magazine articles.
- Like clickbait, *disinformation* images contain many politicians and celebrities, but also include images overlaid with white text.
- *Propaganda* images generally appear as natural scenes.
- Similar to clickbait and disinformation, *satire* images focus on individuals, however not politicians.
- Finally, *verified* images include many natural and unaltered scenes with groups of people.

In the cases where the models give a high confidence score to a tweet, but misclassify it, we see that the images illustrate traits typical of what we know about a class. Because our tweets receive a label based on the news account they retweet, discrepancies may exist between the true label and the assigned class, causing a misclassification.

Agreement across Model Predictions

We examine which tweets cause the most disagreement between the five models, i.e., each class label is predicted at least once for a tweet. In Figure 2, we present examples of

[2] ImageOnly model performance is bounded by the techniques used to extract image vector representations.

[3] To improve 7-way classification model we need to fix for class imbalance using undersampling and oversampling techniques.

[4] Demo video: https://bit.ly/2YVWo5Y

[5] There are many papers on visualization and visual analytics to support data scientists to build and debug machine learning models. A complete synthesis of these works is outside the scope of this paper, but interested readers should consult (Hohman et al. 2019; Cai, Jongejan, and Holbrook 2019).

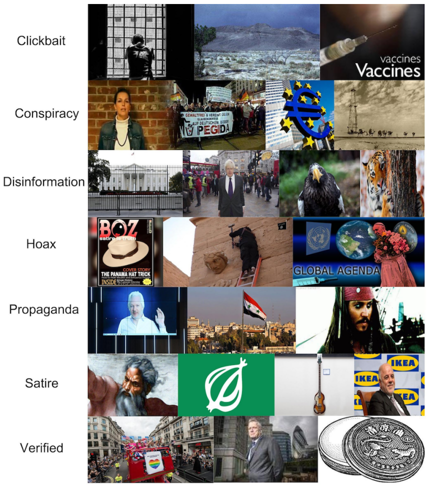


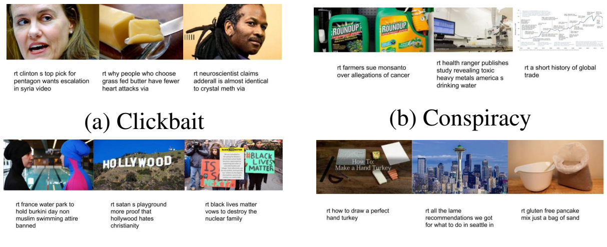
Figure 2: Example images from each deceptive class where none of the models could agree on the label.

images from such tweets. In most cases, we see that these images deviate from our understanding of what characterizes the class in which they belong.

We next look at tweets where all the models assign the same wrong label, e.g. the ground truth is disinformation and all models agreed on propaganda. The class with the highest incorrect prediction in this manner is disinformation (40.08% of tweets) followed by conspiracy (39.13%) and propaganda (37.45%). The least incorrectly predicted class is satire (0.72%), then hoax (2.19%), verified (5.55%), and clickbait (11.26%). Between all collections, about 31.5% of tweets fool all of our models in this way.

Benefits of Multimodal Model Predictions

Based on our results, a combined model of all signals e.g., text and images surpasses individual signal models. Figure 3 illustrates tweets for which isolated text and image features cannot be used for accurate classification, but when joined, pick up on signals that boost performance. We see that this is most helpful for images or text that deviate from the stereotypical example for a particular class.



rt clinton s top pick for pentagon wants escalation in syria video

rt why people who choose grass fed butter have fewer heart attacks via

rt neuroscientist claims adderall is almost identical to crystal meth via

(a) Clickbait

rt farmers sue monsanto over allegations of cancer

rt health ranger publishes study revealing toxic heavy metals america s drinking water

rt a short history of global trade

(b) Conspiracy

rt france water park to hold burkini day non muslim swimming attire banned

rt satan s playground more proof that hollywood hates christianity

rt black lives matter vows to destroy the nuclear family

(c) Hoax

rt how to draw a perfect hand turkey

rt all the lame recommendations we got for what to do in seattle in

rt gluten free pancake mix just a bag of sand

(d) Satire

Figure 3: Clickbait, conspiracy, hoax, and satire tweets correctly classified by the joint Text + (Lex + Image) model, but incorrectly classified by the individual TextOnly and ImageOnly models.

Conclusions

We presented qualitative and quantitative evaluation of multimodal deceptive news prediction models. We contrasted the performance of five neural network models, leveraging combinations of text, lexical, and image signals on four deceptive news prediction tasks. We then performed quantitative analysis using a novel interactive tool, ErrFilter, that revealed characteristic input signals and patterns of errors made by the various prediction models, with the goal to inform future deception prediction models.

References

Cai, C. J.; Jongejan, J.; and Holbrook, J. 2019. The effects of example-based explanations in a machine learning interface. In *IUI*, 258–262.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hohman, F. M.; Kahng, M.; Pienta, R.; and Chau, D. H. 2019. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on VCG*.

Karduni, A.; Cho, I.; Wesslen, R.; Santhanam, S.; Volkova, S.; Arendt, D. L.; Shaikh, S.; and Dou, W. 2019. Vulnerable to misinformation?: Verifi! In *IUI*, 312–323.

Kasra, M.; Shen, C.; and O'Brien, J. F. 2018. Seeing is believing: How people fail to identify fake images on the web. In *Extended Abstracts of CHI*, LBW516.

Kim, J.; Tabibian, B.; Oh, A.; Schölkopf, B.; and Gomez-Rodriguez, M. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *WSDM*, 324–332.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. 2014. Multimodal neural language models. In *ICML*, 595–603.

Lazer, D. M. J., et al. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 1532–1543.

Rashkin, H.; Choi, E.; Jang, J. Y.; Volkova, S.; and Choi, Y. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of EMNLP*, 2931–2937.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*, 230–239.

Volkova, S., and Jang, J. Y. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *The Web Conference*, 575–583.

Volkova, S.; Shaffer, K.; Jang, J. Y.; and Hodas, N. O. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *ACL*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.