Verified Tweets

AJ Erdaty, Rachael Sweetman,
Rebecca O’Donovan

Research Question

Can we use a Twitter user's tweets to predict whether their account is verified or not?

Why be verified in the first place?

- Credibility
  - Trust matters in the media
- Added security
- Shows you are one of the top in your field
  - You have public interest

The Problem

- Misinformation is a prominent part of everyday news
  - 23% of adults have shared fake news
  - Studies show people cannot decipher truthful information and propaganda
  - Respondents believed the fake headlines were “somewhat” or “very” accurate 75% of the time.
- Fake Twitter accounts get confused with real ones very easily

Let's compare...

Donald J. Trump
@realDonaldTrump

Angelina was right to dump Brad. He is a loser. Delusional like Benjamin Button.

Reply Retweet Favorite Her

Donald J. Trump
@realDonaldTrump

I have never seen a thin person drinking Diet Coke

Reply Retweet Favorite Her

...which one is real?

What we did

- Made our own twitter dataset
- Built 4 predictive models on it
  - Baseline
  - Naive Bayes
  - Perceptron
  - CNN
    - Features: Word2Vec trained on Twitter

The Data

Collected 2-10 tweets from 1,000 verified twitter users and 1,000 unverified twitter users

Non-Verified User Selection

Pulled 5,000 most recent followers of the 10 most followed accounts, selected 1,000 randomly from this pool

Verified User Selection

The twitter account @verified follows all verified accounts. Pulled random followers from this account until we reached 1,000

Our Results

Baseline

- Counts the number of each word in the training set in both unverified and verified tweets.
- Observes whether that word was found more often in verified or unverified tweets.
- Add 1 to the ‘value’ of that tweet for each word that was found more often in verified tweets or subtract 1 if found more often in unverified tweets.
- 5 trials: 0.670, 0.663, 0.659, 0.651, 0.691
- Avg accuracy = 66.7%

Naive Bayes

- 5 trials: 0.646, 0.663, 0.714, 0.686, 0.708
- Avg accuracy = 68.3% accuracy

Perceptron

- 5 trials & 10 iterations: 0.773, 0.766, 0.777, 0.769, 0.779
- Avg accuracy = 77.3% accuracy

CNN

- 5 trials: 0.741, 0.735, 0.718, 0.735, 0.736
- Avg accuracy = 73.3% accuracy

What This Means

Proof of concept

- Tying credibility to word usage
  - Right now credibility is clearly signified by Twitter's verified stamp, but our results suggest it's possible that credibility information is also contained in the content of tweets
- Shows the kind of information we are able to extract from tweets
  - This research can be used as a starting point for more in-depth research

Model enhancements / possible future uses

- Overall Validity Prediction
  - A company account that's unverified, but still the official company account
  - Unverified police departments tweeting about the Jersey shooting on Tuesday
    - program to verify if they are the actual police department or a fake

Questions?

Works Cited

- https://www.forbes.com/sites/brettedkins/2016/12/20/americans-believe-they-can-detect-fake-news-studies-show-they-cant/#2b8716b34022
- https://www.forbes.com/sites/tomward/2017/04/14/how-i-got-verified-on-twitter-and-how-you-can-too/#1edd84086493
- https://journalistsresource.org/studies/society/internet/fake-news-conspiracy-theories-journalism-research/
- https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

- Link to our GitHub repo: https://github.com/brAInPower2/twitterverification