FauxBuster: A Content-free Fauxtography Detector Using Social Media Comments

Daniel Zhang[1], Lanyu Shang[1], Biao Geng[1], Shuyue Lai[1], Ke Li[1], Hongmin Zhu[1], Tanvir Amin[2], Dong Wang[1]

[1]Department of Computer Science and Engineering

University of Notre Dame, USA

[2]Google Inc. Mountain View, CA, USA

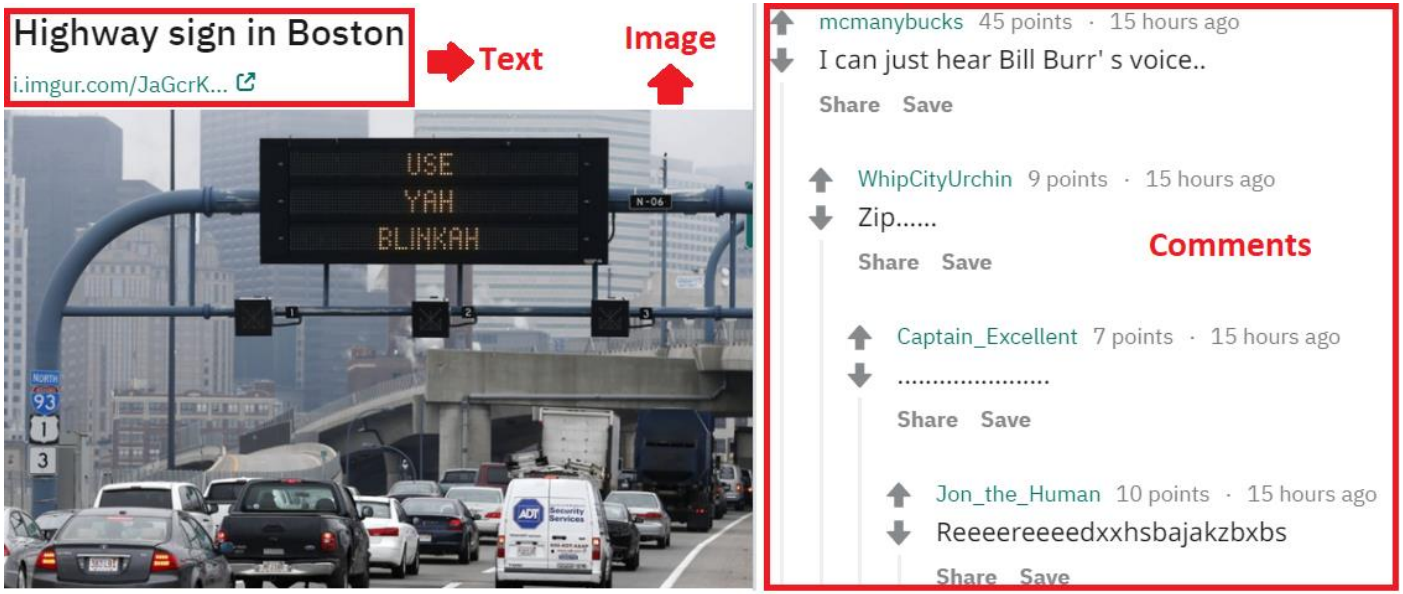
UNIVERSITY OF NOTRE DAME

Google Research

IEEE BigData 2018, Seattle, WA, USA

Motivation: Image-based Social Media Posts

- Image-based posts have become one of the most popular type of content on social media platforms.
- On Twitter, tweets with images get 18% more clicks, 89% more likes, and 150% more retweets than tweets without images.
- On Facebook, photos are found to be the most engaging type of content where 87% of the posted photos have been clicked, liked or shared.



Motivation: the Fauxtography Problem

Definition: the image(s) and the associated text of a social media post that convey a questionable or outright false sense of the events it seems to depict.

The Fauxtography Problem – Both Image and Text Matter

Definition: the image(s) and the associated text of a social media post that convey a questionable or outright false sense of the events it seems to depict.

"Shark Attacking Helicopter" (Fauxtography)

"I think this might be photoshopped" (non-Fauxtography)

Why It's Challenging?

Both real image and real text can lead to fauxtography!

A . Fake Image, Fake Claim

B . Fake Image, True Claim

C . Real Image, Fake Claim

D . Real Image, Real Claim

- A – Putin is pulling Obama's tie (Fake)
- B – sea creatures are falling from the sky in China during a tornado event (True)
- C – a guy was waiting at the finish line to propose to his girlfriend who died during the Boston Marathon Bombing event (Fake)
- D – a wildfire is happening in Tennessee (True)

Existing Solutions

1. Truth Discovery/ Fact Checking
    - **Only detects false claims in textual content**

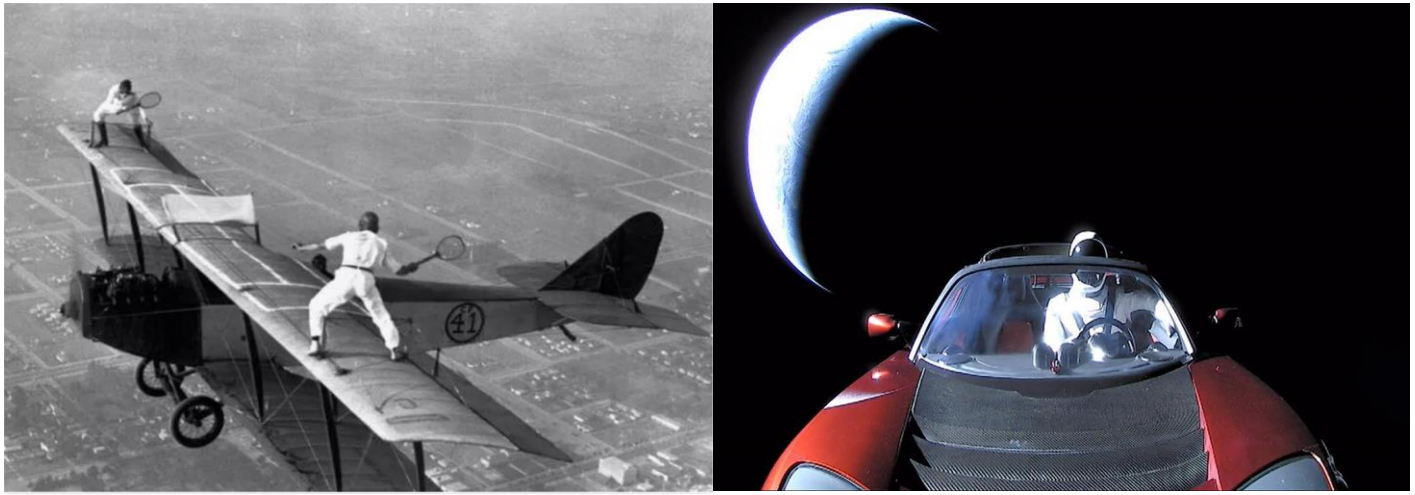
2. Fake Image Detection
    - **Only detects image forgery**

We found no existing solutions can solve the problem where **both image and the context (i.e., the text content)** are taking into account.

How about crowdsourcing solutions?

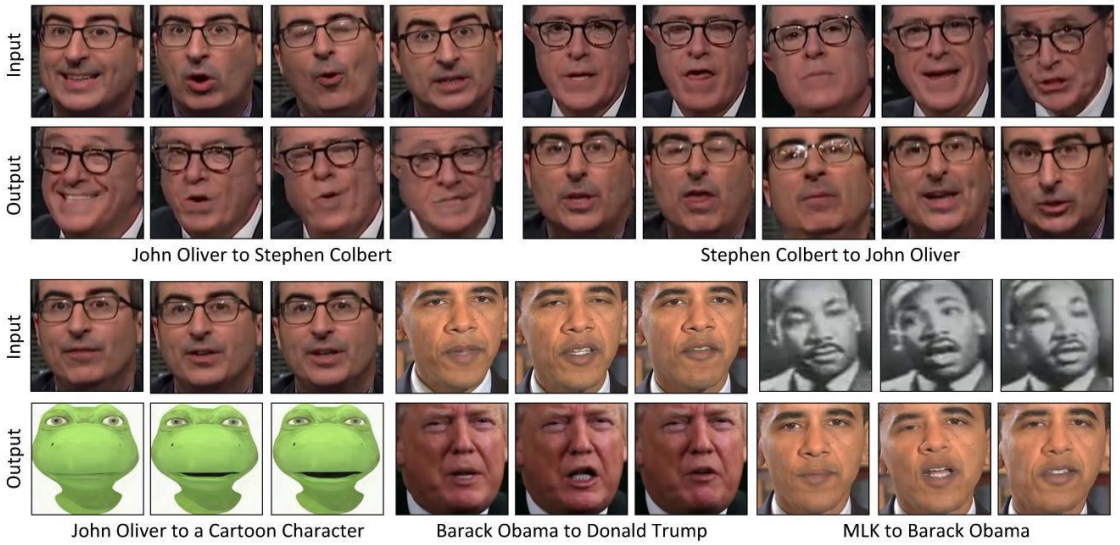
Human actually performed poorly in detecting fauxtography.

Real **Real**



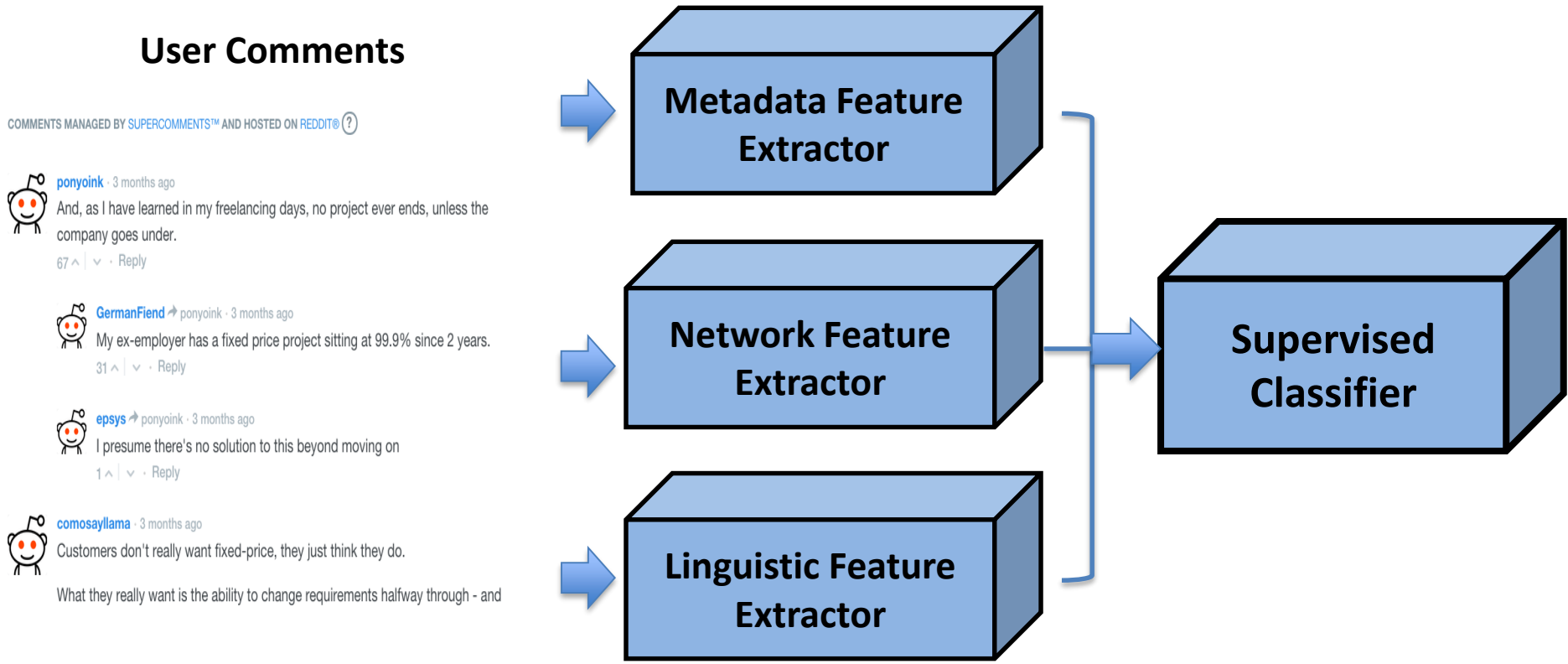
Deep learning algorithms can create real looking images that are hard to distinguish.

Example of "Deep fakes"



Solution Intuition of FauxBuster

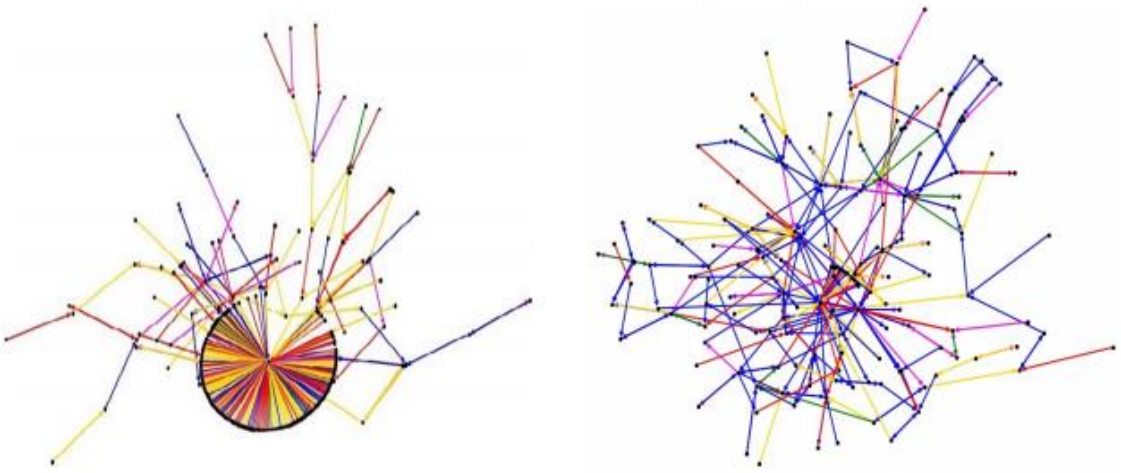
- We develop a supervised learning solution - **FauxBuster** that leverages three types of features from the comments: **network**, **linguistic** and **metadata.**
- FauxBuster is **content-free**: **use the user comments** of a image-based post, **without analyzing the actual image content**.



Network Feature Extractor 1/3

- We first construct three comment networks for each post, where each node represents a comment and each edge represents a directly reply
- For each network, we assign different types of semantic labels (emotion score, attitude score, and feedback score) for the edge

Emotion Network captures the comments' sentiments (**IBM Watson Tone Analyzer)**

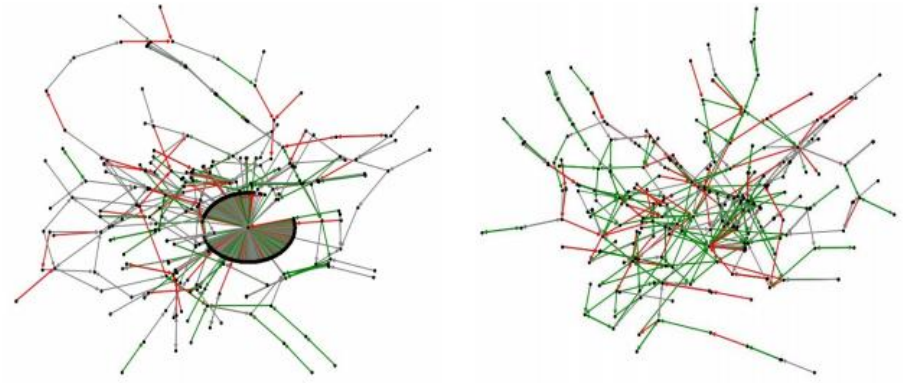


(a) Emotions of Fauxtography (b) Emotions of Non-Fauxtography

Figure : Illustration of emotion features. We use colors to denote the dominant emotion of each comment - "yellow-joy, red-anger, pink-disgust, green-fear, blue-sadness".

Network Feature Extractor 2/3

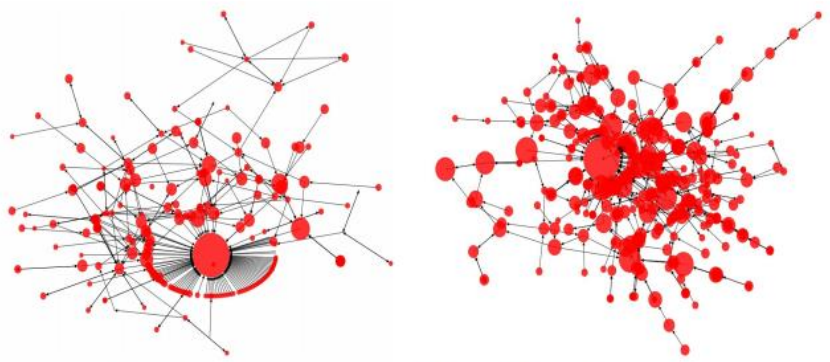
Attitude Network captures the approval/debunking of a comment towards another



(a) Attitudes of Fauxtography (b) Attitudes of Non-Fauxtography

Figure: Illustration of attitude features. We use colors to denote the attitude of each comment - "red-debunk, green-endorse, black-neutral".

Feedback Network captures the likes/dislikes of a comment

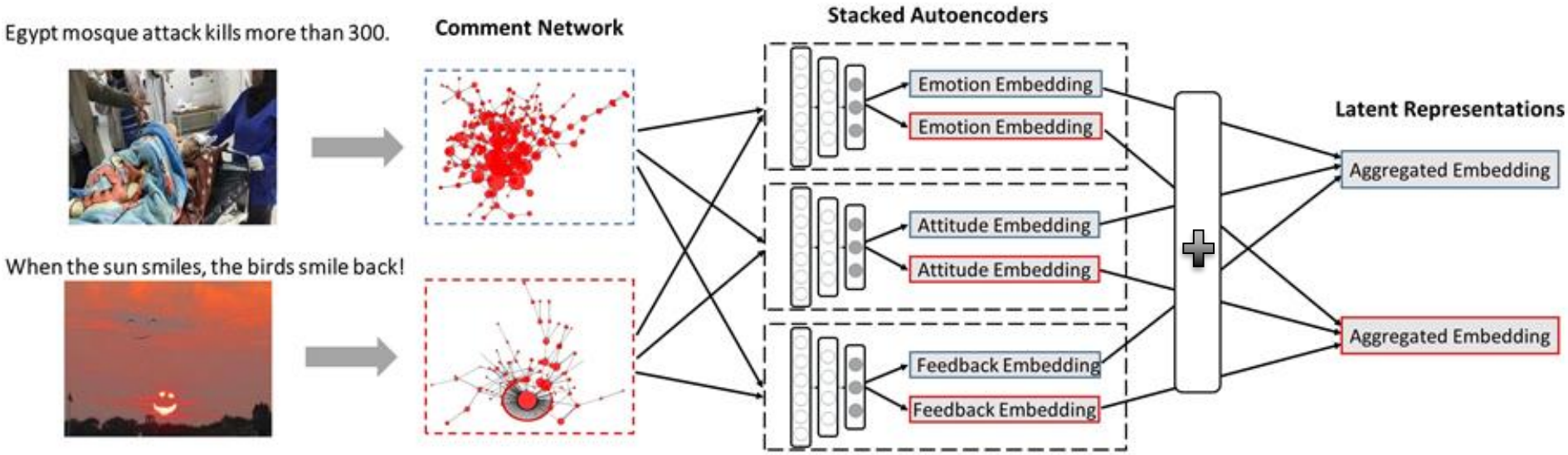


(a) Feedback of Fauxtography (b) Feedback of Non-Fauxtography

Figure: Illustration of feedback features. The size of a vertice indicates the aggregated feedback score (i.e., # of likes - # of dislikes) of all comments from a user.

Network Feature Extractor 3/3

- First perform **random walks** to capture the semantic and topological features of each network
- Each graph is walked M times and each walk has a max length of N – leading to M*N feature vector for each network
- Use **stacked autoencoder** to embed the M*N feature vector



Linguistic Feature Extractor

- Non-fauxtography comments are more related to topic itself
- Fauxtography posts have trolls, funny comments and contains more image related words
- Uses **Doc2Vec** to capture the linguistic feature of the comments

(a) Fauxtography (b) Non-Fauxtography

Figure : Word Cloud

Metadata Feature Extractor

Other features that might be important:

Table : Metadata Features

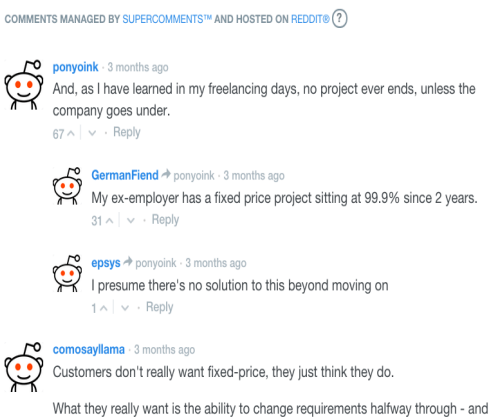
Feature	Description
total comments	Total # of comments in each post
average comments	Avg. # of comments under each thread (Reddit)
average verity	Avg. # of verity related words in each comment
average image	Avg. # of image related words in each comment
average question	Avg. # of question marks in each comment
average exclamation	Avg. # of exclamation mark in each comment
total url	Total # of URLs
average url	Avg. # of comments contain URLs
average word count	Avg. # of words in each comment

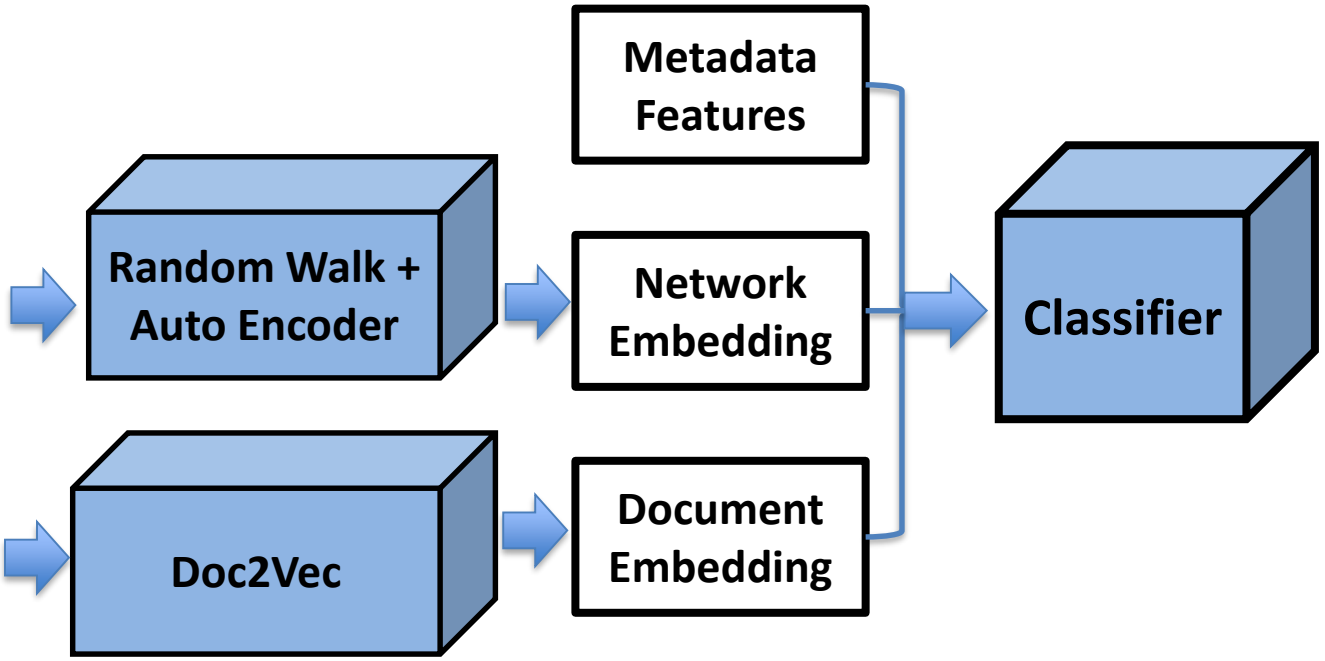
Supervised Classifier

We use existing classifiers:

- **XGBoost**
- **Naïve Bayes (NB)**
- **Random Forest (RF)**
- **Linear Support Vector Machine (SVM)**
- **Multi-layer Perceptron (MLP)**

User Comments





Dataset

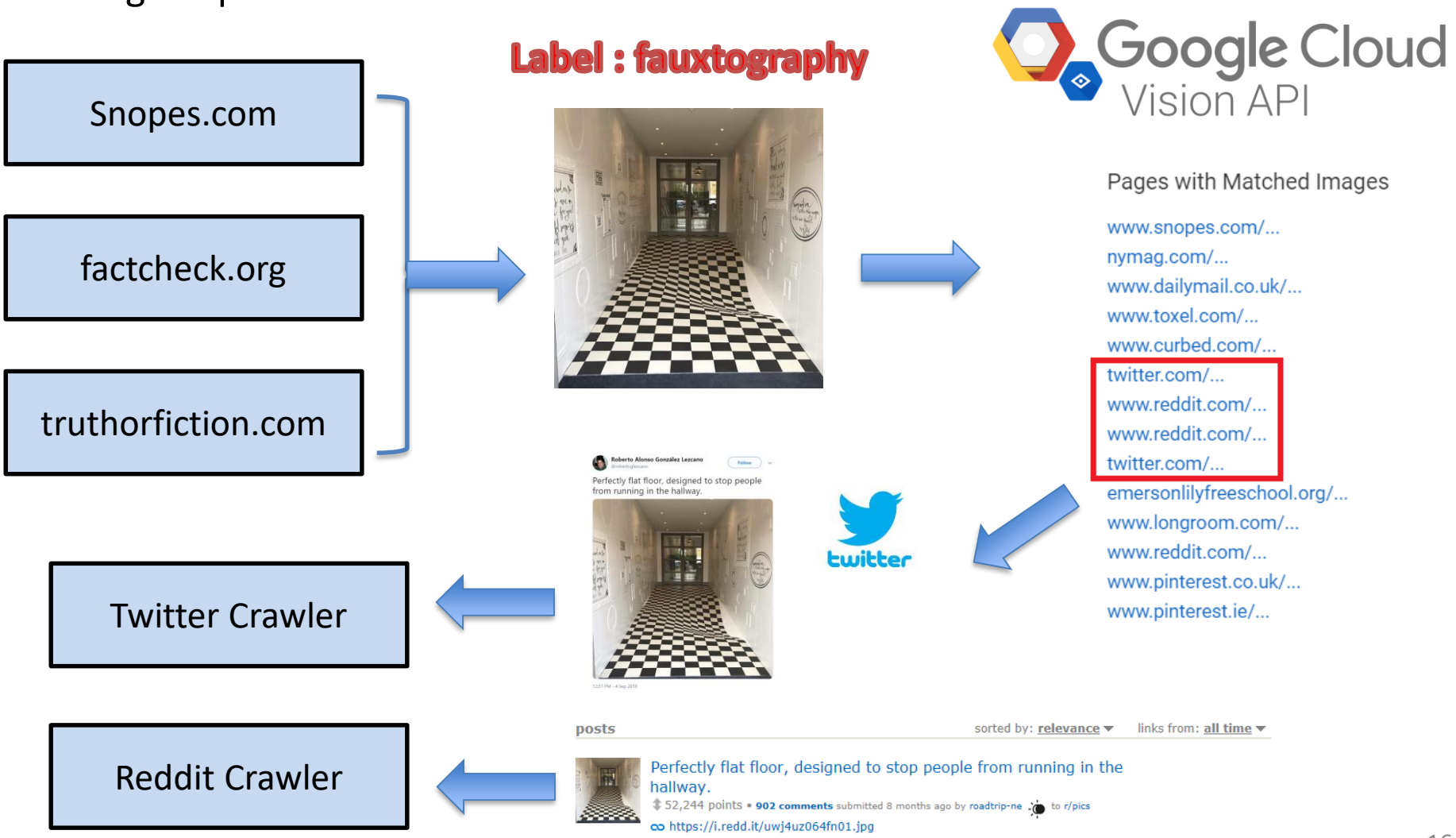
- Two datasets of popular social media sites – Reddit and Twitter
- 196 posts for Reddit, 721 posts for Twitter
- 70% training, 30% testing

Table : Data Trace Statistics

Data Trace	Reddit	Twitter
Number of Posts	196	721
Number of Fauxtography	91	390
Number of Fauxtography with Real Images	12	40
Number of Comments	60,168	1,928,325
Number of Distinct Users	39,702	582,281

Data Collection Process

- **Reverse crawling** – identify ground truth labels first and then find the original posts



Evaluation - Detection Effectiveness

Table: Classification Accuracy for All Schemes

	Reddit				Twitter			
Algorithms	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
XGBoost(FauxBuster)	**0.918**	**0.903**	**0.933**	**0.915**	**0.743**	**0.81**	**0.762**	**0.785**
NB	0.747	0.704	0.905	0.792	0.639	0.746	0.631	0.684
RF	0.864	0.897	0.833	0.864	0.647	0.704	0.738	0.721
SVM	0.559	0.559	0.633	0.594	0.5	0.603	0.559	0.58
MLP	0.608	0.6	0.786	0.68	0.647	0.757	0.631	0.688
Fake Image	0.835	0.837	0.857	0.847	0.603	0.75	0.536	0.625
Truth Discovery	0.683	0.73	0.643	0.683	0.529	0.632	0.571	0.601

- XGBoost has consistently better performance
- Significantly outperform baselines, including Fake Image detection and Truth Discovery schemes

Evaluation - Detection Time

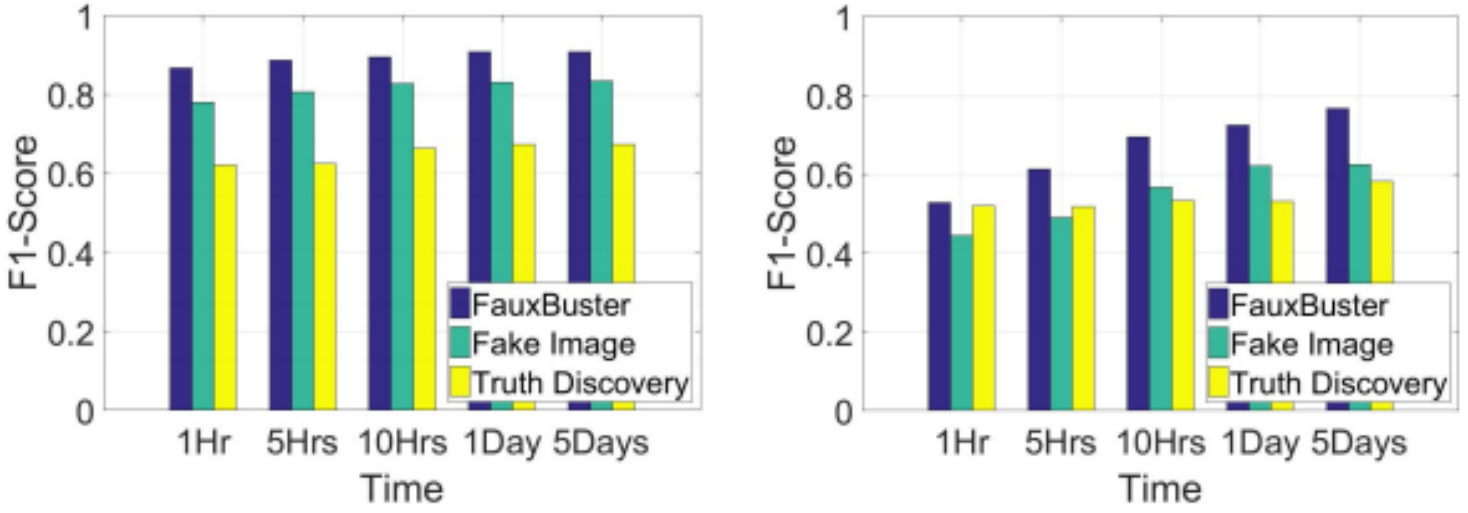


Figure: Elapsed Time vs. Performance

- Consistently outperform baselines at different timeframes.

Evaluation – Feature Analysis

Table: Feature Analysis for FauxBuster

	Reddit				Twitter			
Feature Sets	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
All	**0.918**	**0.903**	**0.933**	**0.915**	**0.743**	**0.81**	**0.762**	**0.785**
Network Only	0.711	0.709	0.762	0.736	0.7	0.759	0.717	0.737
Linguistic Only	0.747	0.806	0.691	0.744	0.684	0.747	0.738	0.742
Metadata Only	0.823	0.937	0.714	0.811	0.566	0.658	0.619	0.638
Network + Linguistic	0.772	0.853	0.691	0.763	0.692	0.735	0.735	0.735
Network + Metadata	0.899	0.925	0.881	0.902	0.654	0.761	0.643	0.697
Linguistic + Metadata	0.899	0.947	0.857	0.9	0.639	0.673	0.735	0.702

- Network features, metadata features, and linguistic features all pay important roles
- All three features together delivers the best performance

Evaluation – FauxBuster vs. Human

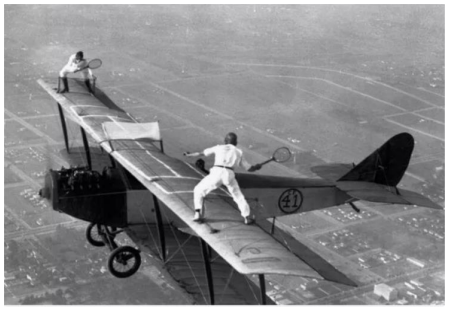
- Three annotators
- First label without comment
- Then label with access to comment

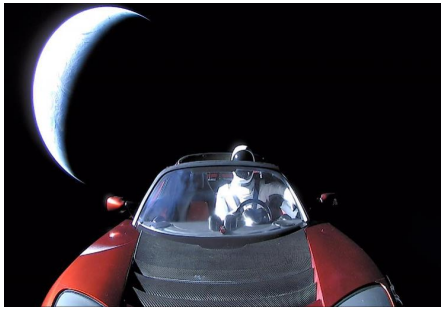
Table: FauxBuster vs. Human Performance

	Accuracy	F1	FPR	FNR
FauxBuster	**0.92**	**0.915**	**0.058**	**0.104**
A1	0.44	0.391	0.422	0.672
A1+comment	0.71	0.713	0.222	0.345
A2	0.46	0.413	0.4	0.654
A2+comment	0.7	0.737	0.378	0.236
A3	0.39	0.408	0.6	0.618
A3+comment	0.63	0.648	0.356	0.382
Overall	0.44	0.404	0.444	0.654
Overall+comment	0.74	0.764	0.289	0.236

* "Overall" denotes the majority vote of the three annotators.

Top False Positives





Top False Negatives

Conclusion and Future Work

- We address the emerging issue of fauxtography in social media posts.
- We developed the first content-free tool for fauxtography detection.
- We studied interesting and unique hints from users' comments.
- We demonstrated that the proposed FauxBuster scheme outperforms existing baselines.

Next Steps

Real-time Solution

- Detect fauxtography on-the-fly

Dynamic Network Analysis

- Study the evolving patterns of comment networks

Q&A

Thank You!


VITA DUL-
CEDO SPES
UNIVERSITY OF
NOTRE DAME


Google
Research


NSF


SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951