Misleading Metadata Detection on YouTube

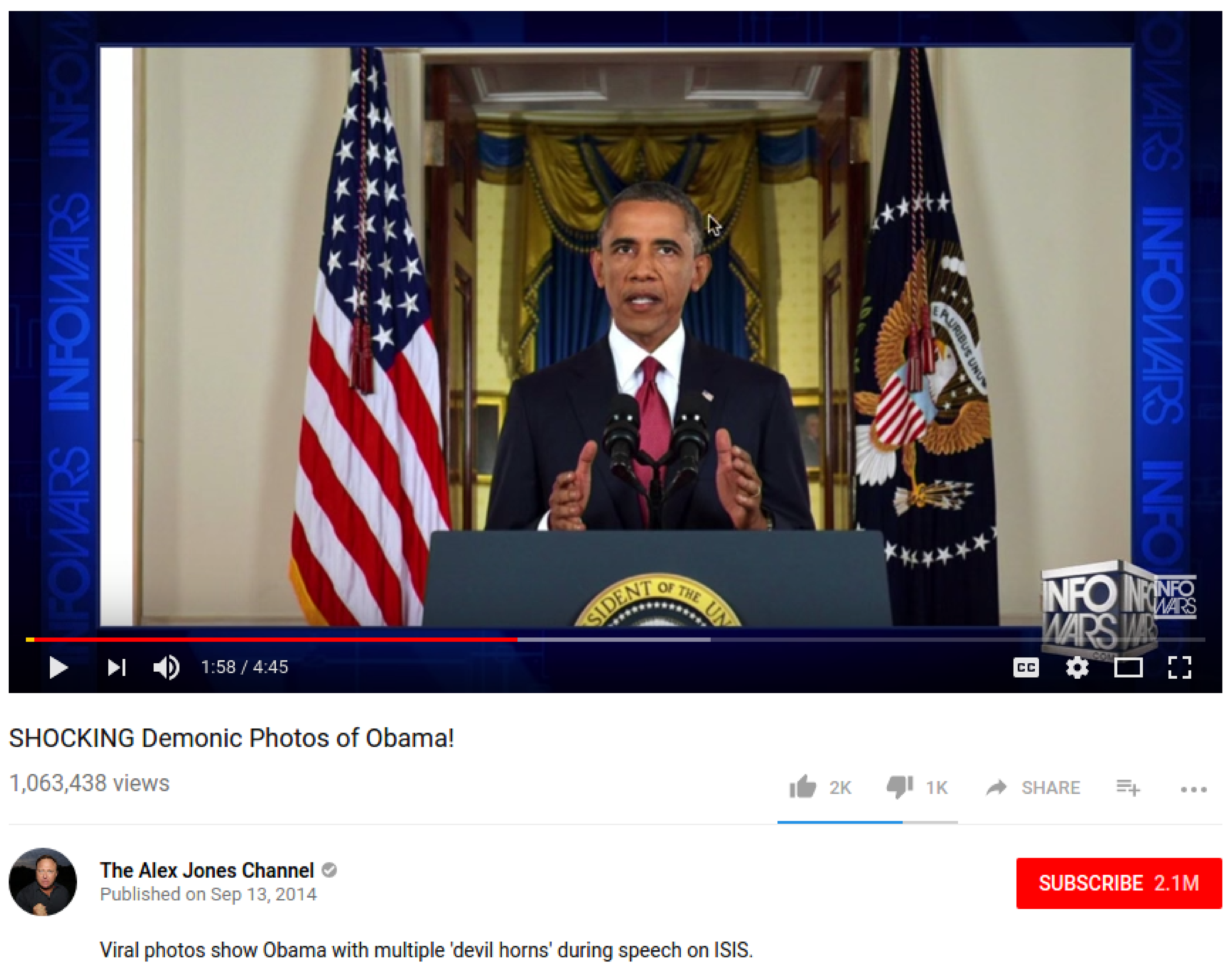
Priyank Palod[1], Ayush Patwari[2], Sudhanshu Bahety[3], Saurabh Bagchi[2], Pawan Goyal[1]

[1]Dept. of CSE, IIT Kharagpur, India; [2]Purdue University; [3]Salesforce.com

1 Objective

YouTube is plagued with misleading content that includes staged videos presented as real footages from an incident, videos with misrepresented context and videos where audio/video content is morphed. We tackle the problem of detecting such misleading videos as a supervised classification task.

2 Example Video



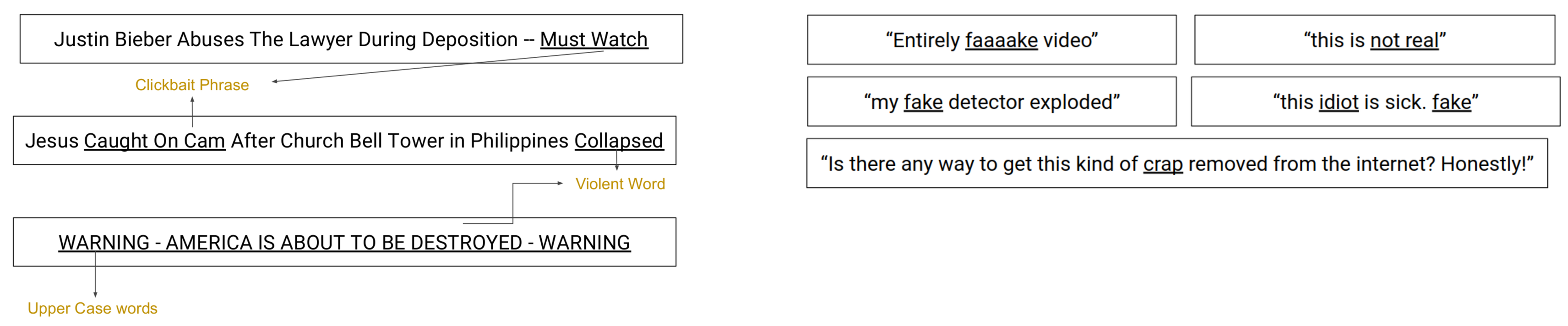
3 Datasets

- **Fake Video Corpus (FVC) [1]**
  - Had 117 fake and 110 real video URLs, but some got removed. Used 98 fake and 72 real videos.
  - The paper reported 79% F-Score(fake class), but we found 36% Macro-Avg.
  - We divided it in 30:70 ratio and called the subsets FVC30 and FVC70 respectively.
- **Volunteer annotated Video Dataset (VAVD)**
  - Crawled 100K video urls from YouTube. Removed videos with views < 10k, comments < 120
  - Handpicked phrases from some fake videos and bootstrapped (e.g., “complete bullshit”)
  - Removed videos with dislike count:like count < 0.3 and got 650 videos to be annotated by students.
  - After annotations: 421 Real videos, 125 Fake videos. 104 videos - not sure (these are ignored).

4 Some Example Simple Features

Title/Description Based: Presence of Clickbait phrase, Ratio of UpperCase:LowerCase words in Title, Ratio violent words in title, etc.

Comment Based: Ratio of comments with swear words, fakeness indicating words.



5 Experiments with Simple Features

Classifier	Precision	Recall	F-Score
SVM- RBF	0.74	0.60	0.49
Random Forests	0.73	0.58	0.46
Logistic Regression	0.54	0.53	0.45
Decision Tree	0.53	0.52	0.46

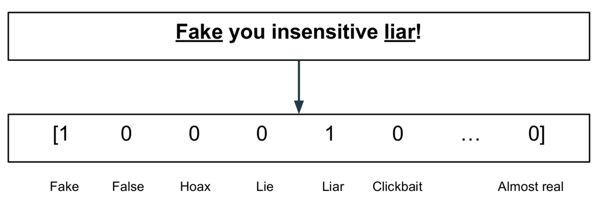
Table 1: Simple classifiers train: VAVD, test:FVC30

Classifier	Precision	Recall	F-Score
SVM- RBF	0.56	0.55	0.54
Random Forests	0.74	0.73	0.73
Logistic Regression	0.53	0.53	0.53
Decision Tree	0.73	0.67	0.67

Table 2: Simple classifiers train: FVC70, test: FVC30

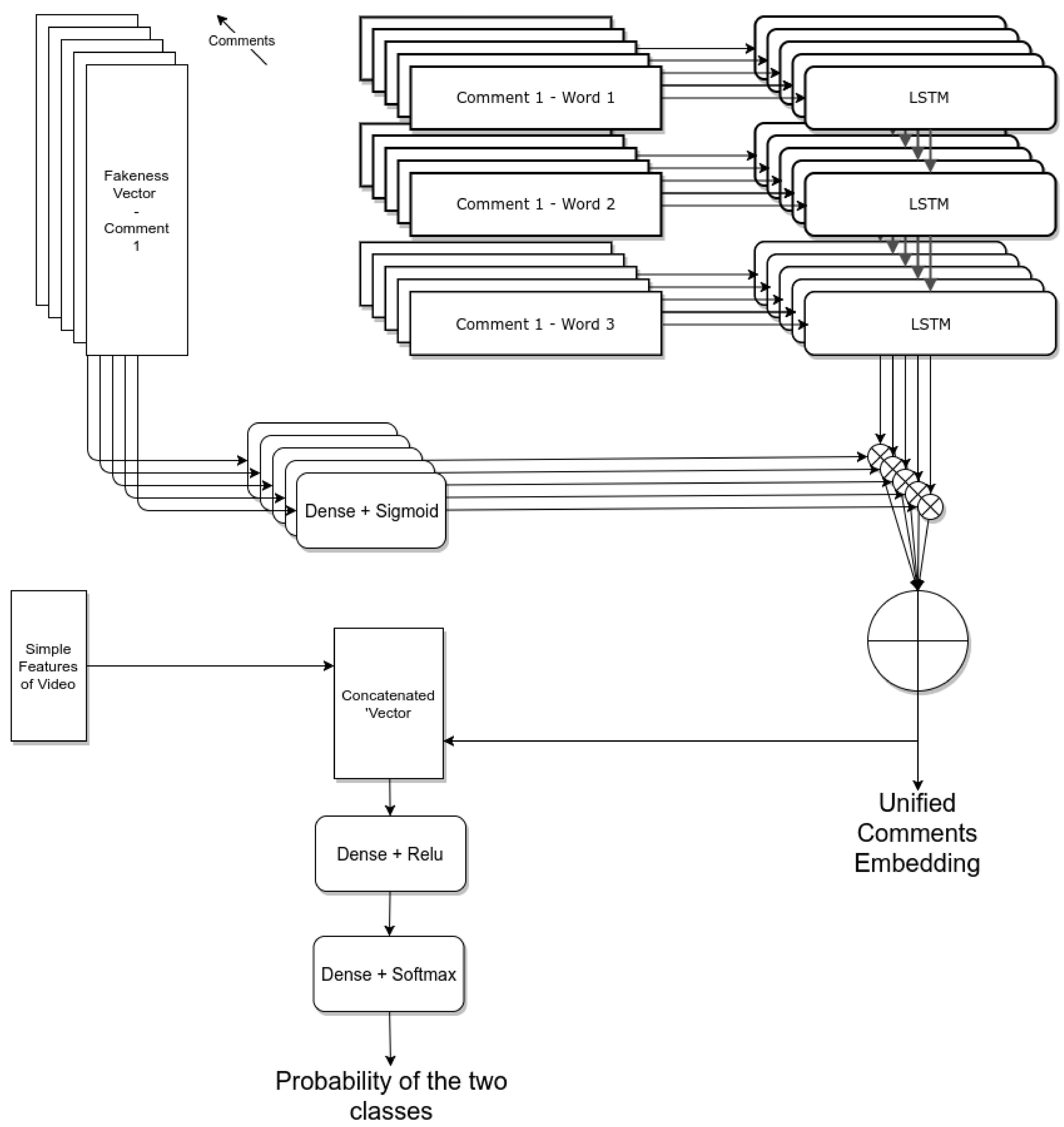
6 UCNet: Description

- Create a “fakeness indicator vector” for each comment using some words/phrases.



- Pass it through a dense layer with sigmoid activation to get a ‘weight’ of the comment (0-1).
- Get an embedding of each comment by passing it word by word (word2vec) through LSTM.
- Take the weighted average of all comments called “Unified Comments Embedding” (UCE).
- Concatenate UCE with Simple features and pass through dense layers for classification.

7 UCNet: Diagram



8 Results with UCNet

Class	Precision	Recall	F-score	#Videos
Real	0.64	0.88	0.74	72
Fake	0.88	0.64	0.74	98
Macro avg	0.76	0.76	0.74	170

Table 3: UCNet train: VAVD, test: FVC

Class	Precision	Recall	F-score	#Videos
Real	0.74	0.87	0.8	23
fake	0.89	0.77	0.83	31
Macro avg	0.82	0.82	0.82	54

Table 4: UCNet: train: FVC70, test: FVC30

9 PCA further demonstrates importance

Figure 1: PCA for Simple Features

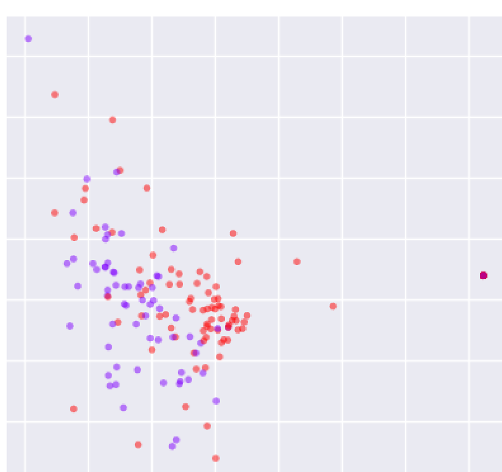


Figure 2: PCA for UCE

- Red dots are Fake Videos while blue dots are Real Videos
- UCE can distinguish between the fake and real videos better than the simple features.

10 Conclusion

Our work presents VAVD, a new dataset for research on fake videos, and also presents UCNet, a deep learning based approach to identify fake videos with high accuracy using user comments. UCNet also generalizes well across datasets.

- **Dataset**: https://github.com/ucnet01/Annotations_UCNet
- **Code:** https://github.com/ucnet01/UCNet_Implementation

11 References

[1] Papadopoulou, O., Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Web video verification using contextual cues. In: 2nd Intl. Workshop Multimedia Forensics and Security. pp. 6–10. ACM (2017)