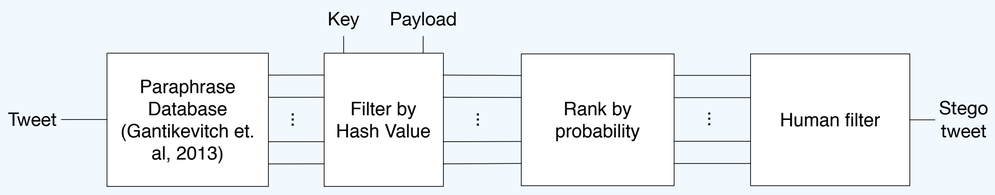
Detecting Steganographic Techniques on Twitter

Alex Wilson, Phil Blunsom and Andrew D Ker
Department of Computer Science
University of Oxford

Linguistic Steganography and CoverTweet

Steganography hides information in natural language.
CoverTweet uses automatic paraphrasing to hide data in Tweets:



The system finds a paraphrase with a desired hash value.
The goal is to hide as much data as possible, while avoiding detection.
We showed that tweets containing 4 bits were undetectable to human judges (Wilson et al., 2014).

Which two of these do you think are hiding information?

1. `if anybody wanted to text me i wouldn't even mind.`
2. `i should probably go to sleep soon.`
3. `sleep's overrated..`
4. `this summer is going to be the best one yet.`

`hash value = 0110`

(Answer: one and three)

Pooled Steganalysis

Individual stego tweets are difficult to detect.
Pooled steganalysis is an alternative paradigm, where we attempt to identify *users* of steganography.
We pool together evidence from multiple tweets by the same user.

Which user do you think is hiding information?

User A

1. `i didn't even get tan. i was sunburnt. adorable`
2. `i've got an a in gym.`
3. `you just do not even realise`
4. `i really hate getting ready`

User B

1. `time to go make some friends`
2. `i think i need to go to the hospital`
3. `i better go sleep`
4. `he is forever eating, i swear`

For each feature, we took the minimum, maximum, mean and variance over batches of tweets.

We trained the classifier on these pooled features.

(Answer: User A)

Data

- Used CoverTweet to hide data in 1000 tweets (100 from 10 users) taken from the Harvard TweetMap (Mostak, 2013).
- Generating stego data with a human filter is expensive, so we also automated the generation of 1M stego tweets (1000 from 1000 users).
- For the automatic data, the paraphrased tweet with the highest probability was selected.
- For comparison purposes, we also hid data in approx. 250k tweets using T-Lex, an older stegosystem.

Statistical Detection

Trained a linear classifier ensemble on four sets of features:

- Basic tweet statistics (e.g. word count, stop word usage)
- n-gram language model probabilities
- word length model probabilities
- features derived from the Paraphrase Database (e.g. estimated probability of phrases in the tweet having been changed by the database)

(This last set of features is allowed under Kerckhoffs' Principle.)

Kerckhoffs' principle

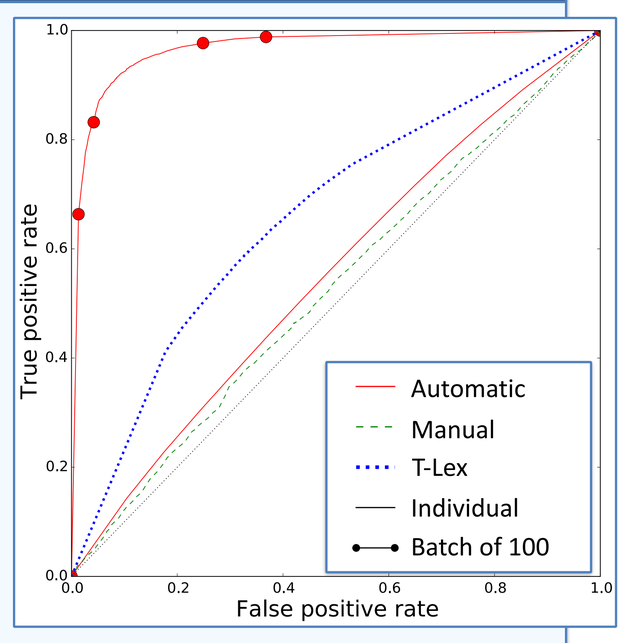
We must assume the attacker of a system knows exactly how the system works.
This means that the attacker has the source of paraphrase rules used by the steganographer.

Results

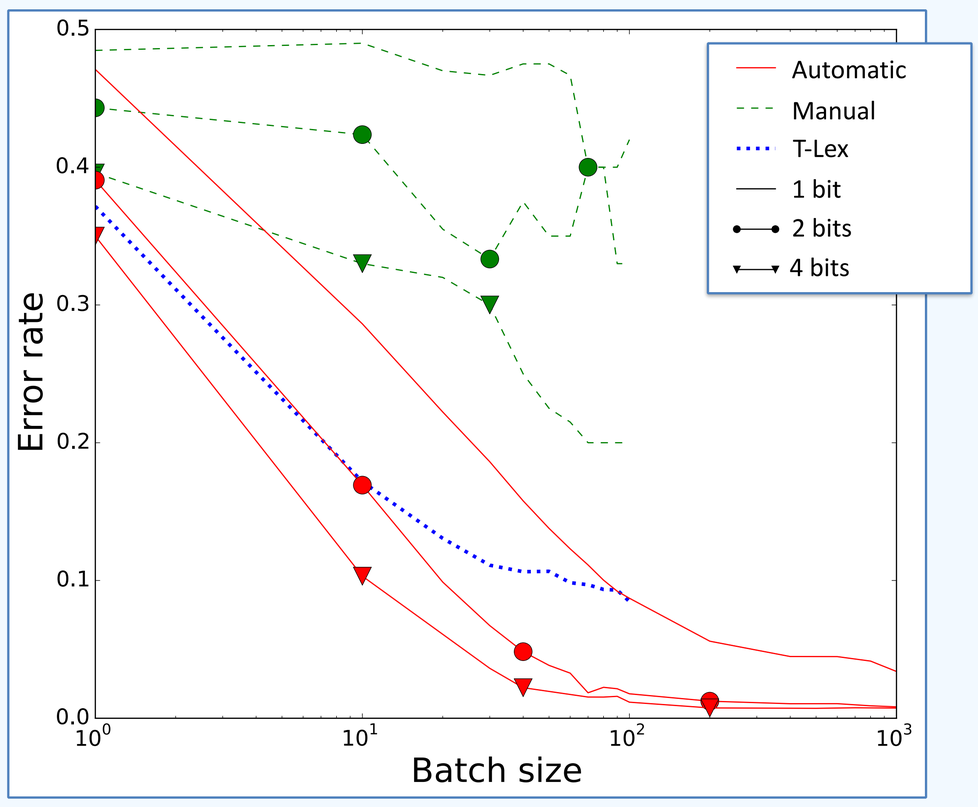
The performance of the trained classifier on individual tweets is poor.

	AUC
Automatic	0.551
Manual	0.509
T-Lex	0.667
Automatic (pooled)	0.963

Pooling the data for each user provided a vast improvement to error rate.



Increasing the batch size (the number of tweets pooled) further reduced the error rate.
Additionally, hiding more data made the tweets easier to spot.



Evaluating each feature set individually showed that those derived from the Paraphrase Database were strongest.

Main Findings

- Individual steganographic tweets are hard to detect.
- Looking at multiple tweets at once allows us to spot the users who are hiding information.
- Knowing the details of the system gives the attacker a powerful advantage.

Contact details: alex.wilson@cs.ox.ac.uk