Stay On-Topic: Generating Context-specific Fake Restaurant Reviews

Automated crowd-turfing

- Puppet master controls sock puppet accounts
- Puppet master dispatches positive / negative reviews to target to influence public opinion
- Advances in natural language processing → AI-written reviews, humans not needed
- Previous approach, LSTM-Fake [1], cannot maintain context in reviews → detectable

Our approach: NMT-Fake*

- Libraries for neural machine translation (NMT) explicitly condition text generation on context
- Adversary can use these libraries to generate context-specific reviews? E.g.

  Context: 5 Gina’s Place Cleveland OH
  Diners Breakfast

  Generated review:

  "The best breakfast place in Cleveland. Great prices and great service. I highly recommend the homemade eggs Benedict. It's a must try!"

- Generation controlled with two parameters: 
  B ~ proportion of novel words
  L ~ importance of using only novel words

User study

- Amazon mTurkers presented with 50% fake and 50% real reviews → parameter search for (L,B).

<table>
<thead>
<tr>
<th>Native English mTurkers detection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
</tr>
<tr>
<td>45%</td>
</tr>
</tbody>
</table>

- Some parameter combinations more detectable. Best combination (II) detected only 68/171 times
- Skeptical user study with expert participants, conditioned to fake reviews. Task: detect 4 machine-written reviews among 30 reviews.
- Skeptical users
  - as good as random: on average 0.8/4 NMT-Fakes* detected
  - statistically worse at detecting NMT-Fakes* than LSTM-Fakes [1] (99% confidence)

How to deal with NMT-Fake*?

- Short term solution: AdaBoost-based detection: 97% effectiveness (macro F1-score)
- Generalizability to other application areas?
- Risk of releasing large textual datasets?