Generating Natural Language Adversarial Examples (Supplementary Material)

Moustafa Alzantot[1*], Yash Sharma[2*], Ahmed Elgohary[3],
Bo-Jhang Ho[1], Mani B. Srivastava[1], Kai-Wei Chang[1]

[1]Department of Computer Science, University of California, Los Angeles (UCLA)
{malzantot, bojhang, mbs, kwchang}@ucla.edu
[2]Cooper Union sharma2@cooper.edu
[3]Computer Science Department, University of Maryland elgohary@cs.umd.edu

1 Additional Sentiment Analysis Results

On the following page(s), Table 1 shows an additional set of attack results against the sentiment analysis model described in our paper.

2 Additional Textual Entailment Results

On the following page(s), Table 2 shows an additional set of attack results against the textual entailment model described in our paper.

*Moustafa Alzantot and Yash Sharma contribute equally to this work.

Original Text Prediction = **Positive**. (Confidence = 78%)
The promise of Martin Donovan playing Jesus was, **quite honestly**, enough to get me to see the film. Definitely worthwhile; clever and funny without overdoing it. The low quality filming was **probably** an appropriate effect but ended up being a little too jarring, and the ending sounded more like a PBS program than Hartley. Still, too many memorable lines and great moments for me to judge it harshly.
Adversarial Text Prediction = **Negative**. (Confidence = 59.9%)
The promise of Martin Donovan playing Jesus was, **utterly frankly**, enough to get me to see the film. Definitely worthwhile; clever and funny without overdoing it. The low quality filming was **presumably** an appropriate effect but ended up being a little too jarring, and the ending sounded more like a PBS program than Hartley. Still, too many memorable lines and great moments for me to judge it harshly.

Original Text Prediction = **Negative**. (Confidence = 74.30%)
Some **sort** of accolade must be given to 'Hellraiser: Bloodline'. It's **actually** out full-mooned Full Moon. It bears all the marks of, say, your 'demonic toys' or 'puppet master' series, without their **dopey**, uh, charm? Full Moon can get away with **silly** product because they know it's silly. These Hellraiser things, man, do they ever take themselves seriously. This **increasingly stupid** franchise (**though** not nearly as **stupid** as I am for having **watched** it) once made up for its low budgets by being stylish. Now it's just ish.
Adversarial Text Prediction = **Positive**. (Confidence = 51.03%)
Some **kind** of accolade must be given to 'Hellraiser: Bloodline'. it's **truly** out full-mooned Full Moon. It bears all the marks of, say, your 'demonic toys' or 'puppet master' series, without their **silly**, uh, charm? Full Moon can get away with **daft** product because they know it's silly. These Hellraiser things, man, do they ever take themselves seriously. This **steadily daft** franchise (**whilst** not nearly as **daft** as i am for having **witnessed** it) once made up for its low budgets by being stylish. Now it's just ish.

Original Text Prediction = **Negative**. (Confidence = 50.53%)
Thinly-cloaked retelling of the garden-of-eden story – nothing new, nothing shocking, although I feel that is what the filmmakers were going for. The idea is **trite**. Strong performance from Daisy Eagan, that's about it. I believed she was 13, and I was interested in her character, the rest left me cold.
Adversarial Text Prediction = **Positive**. (Confidence = 63.04%)
Thinly-cloaked retelling of the garden-of-eden story – nothing new, nothing shocking, although I feel that is what the filmmakers were going for. The idea is **petty**. Strong performance from Daisy Eagan, that's about it. I believed she was 13, and I was interested in her character, the rest left me cold.

Table 1: Example of attack results against the sentiment analysis model. Modified words are highlighted in green and red for the original and adversarial texts, respectively.

Original Text Prediction: **Contradiction** (Confidence = 91%)
Premise: A man and a woman stand in front of a Christmas tree contemplating a single thought. **Hypothesis:** Two **people talk** loudly in front of a cactus.
Adversarial Text Prediction: **Entailment** (Confidence = 51%)
Premise: A man and a woman stand in front of a Christmas tree contemplating a single thought. **Hypothesis:** Two **humans chitchat** loudly in front of a cactus.

Original Text Prediction: **Contradiction** (Confidence = 94%)
Premise: A young girl wearing yellow shorts and a white tank top using a cane pole to fish at a small pond. **Hypothesis:** A girl wearing a **dress** looks off a **cliff**.
Adversarial Text Prediction: **Entailment** (Confidence = 40%)
Premise: A young girl wearing yellow shorts and a white tank top using a cane pole to fish at a small pond. **Hypothesis:** A girl wearing a **skirt** looks off a **ravine**.

Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: A large group of protesters are walking down the street with signs. **Hypothesis:** Some people are holding up **signs** of protest in the street.
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: A large group of protesters are walking down the street with signs. **Hypothesis:** Some people are holding up **signals** of protest in the street.

Table 2: Example of attack results against the textual entailment model. Modified words are highlighted in green and red for the original and adversarial texts, respectively.