**Fine-Grained SPOILER!! Detection from Large-Scale Review Corpora**

Mengting Wan¹, Rishabh Misra², Ndapa Nakashole¹, Julian McAuley¹

¹University of California, San Diego, USA; ²Twitter, USA

Harry Potter and the Deathly Hallows [review author i.e., user] [review subject i.e., item]

| p=0.35 | This was a perfect, Allan blissfully, end to the series. |
| p=0.81 | Though there were deaths that were definitely unprovoked, |
| p=0.44 | I thought we could have spent more time at Hogwarts, as one of |
| p=0.66 | Harry Potter was a really, really great series that I think will be (and |

**Goodreads Dataset**

**Statistics:**
- 1,378,033 English book reviews;
- Across 25,475 books and 18,892 users from goodreads.com;
- Each book/user has at least one associated spoiler review;
- Include 17,872,965 sentences, 3.22% of which are labeled as 'spoiler sentences.'

**Summary of Insights:**
- **Sentence Dependency:** Spoiler sentences generally tend to appear together in the latter part of a review document.
- **Item/User Spoiler Bias:** Distributions of self-reported spoiler labels are highly skewed indicating significantly different spoiler tendencies across users and items.

**Method: SpoilerNet**

**Figure:** Model architecture of SpoilerNet

**Table:** Spoiler sentence detection results on Goodreads and TV Tropes, where arrows indicate the performance boost (↑) or drop (↓) compared with the base model in each group. Best results are highlighted.

**Experiment**

**Datasets:** Goodreads & TV Tropes (16,261 single-sentence comments about 884 TV programs).

**Baselines:** SVM, SVM-BOW (weighted averages of fasttext word embeddings), CNN, HAN (hierarchical attention network). We add the item-specificity features and the item/user bias respectively on the above baselines; remove each of the word attention module, the pre-trained word embedding initialization, and the sentence encoder from HAN to evaluate their performance.

**Evaluation Metrics:** Due to the possible subjectivity of users’ self-reported spoiler tags, we use the area under the ROC curve (AUC), i.e., we expect a positive spoiler sentence is ranked higher than a negative non-spoiler sentence (within the same document or across the entire corpus).

**Table: Constructed Standards of Spoiler Tags:** Review example from The Hunger Games.

**Error Analysis**

**Table:** Distracted by Revelatory Terms: Review example from Murder on the Orient Express.