

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

Mengting Wan¹, Rishabh Misra², Ndapa Nakashole¹, Julian McAuley¹

{m5wan, r1misra, nnakashole, jmcauley}@ucsd.edu ¹ University of California, San Diego, USA; ² Twitter, USA

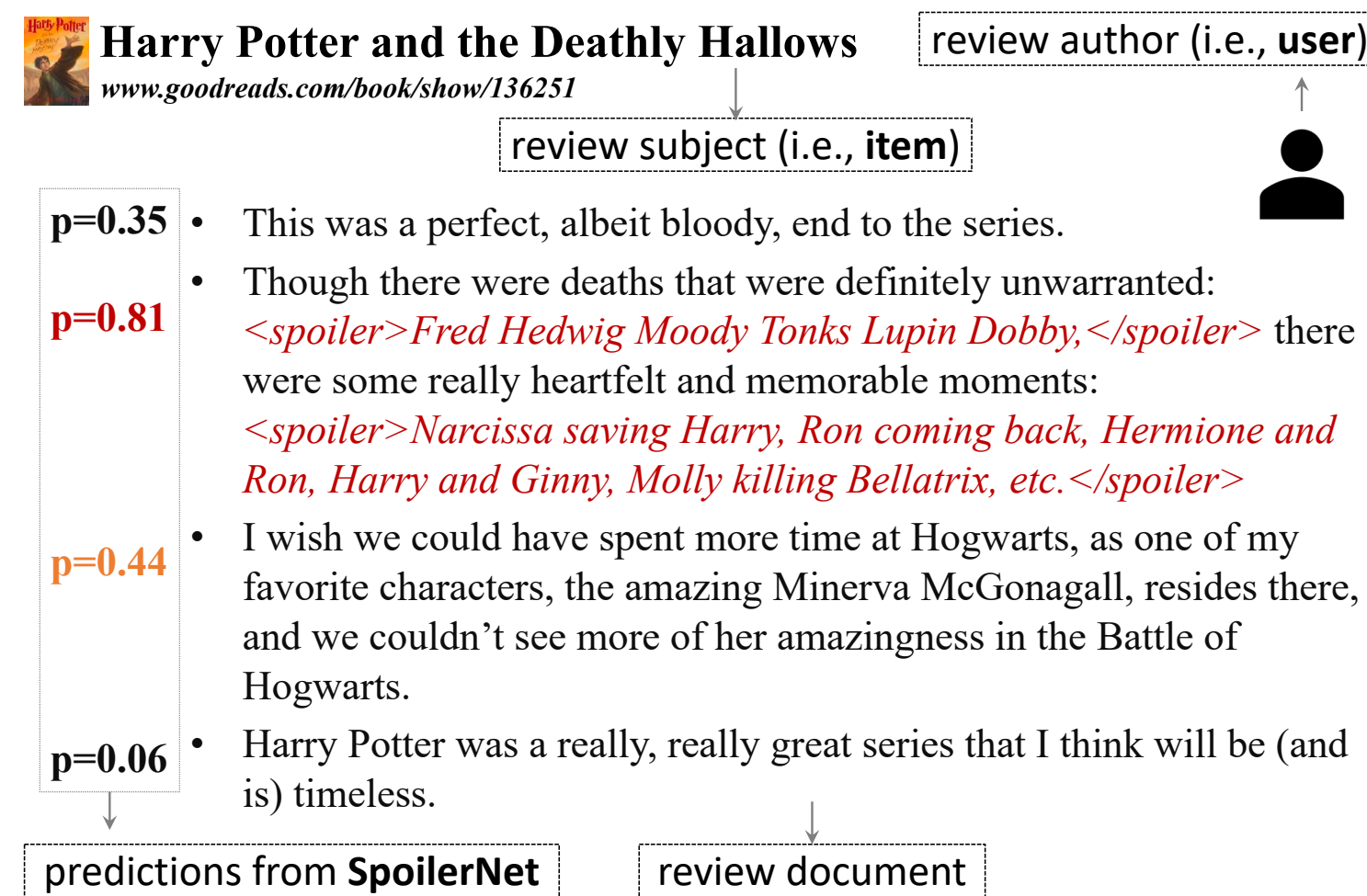


Figure: An example review from *Goodreads*.

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 17,672,655 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.

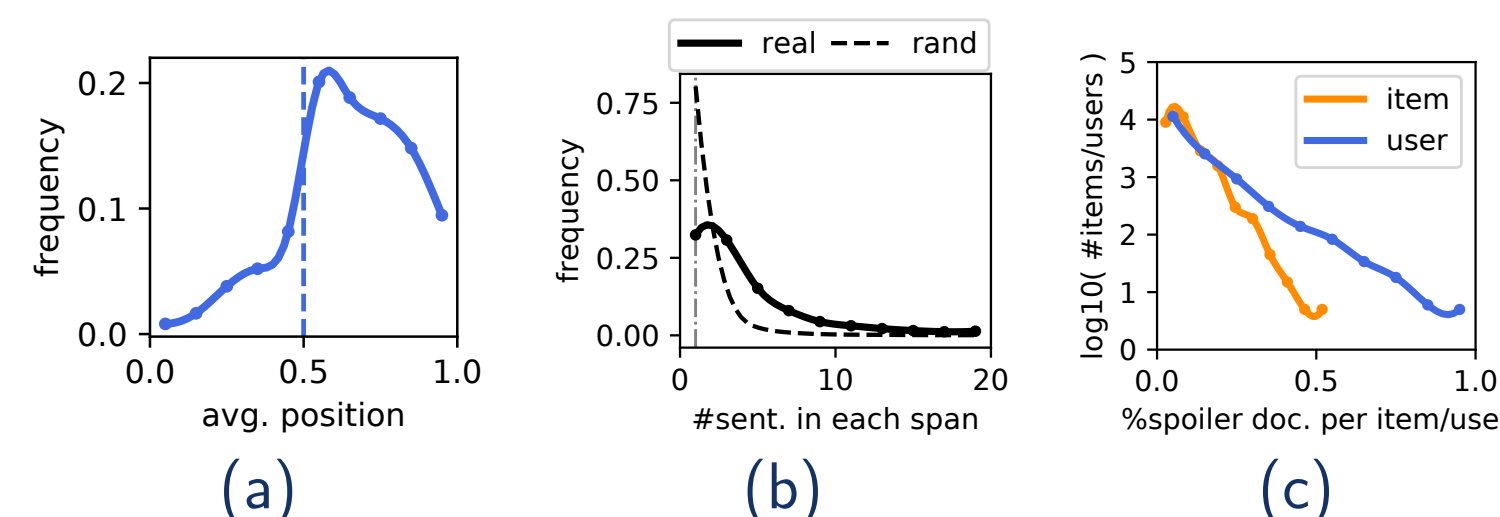


Figure: Distributions of (a) average spoiler sentence position; (b) the length of each spoiler span; (c) the percentage of spoiler reviews per book/user.

Goodreads Dataset (cont'd)

- Item-specificity:** Book-specific terms such as locations or characters' names could be informative to reveal plot information. The item-specificity metric Document-Frequency and Inverse-Item-Frequency ($DF_{w,i} \times IIF_w$) is defined as

$$DF_{w,i} = \frac{|\mathcal{D}_{w,i}|}{|\mathcal{D}_i|}, \quad IIF_w = \log \frac{|\mathcal{I}| + \epsilon}{|\mathcal{I}_w| + \epsilon}$$

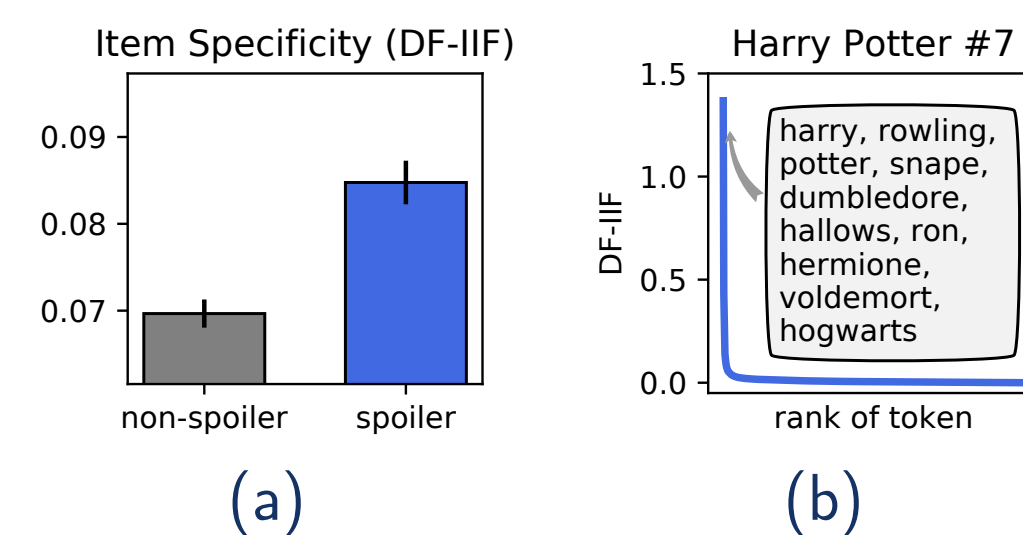


Figure: Distributions of (a) item-specificity of non-spoiler and spoiler sentences; (b) DF-IIF of each term and top ranked item-specific terms for an example book.

Method: SpoilerNet

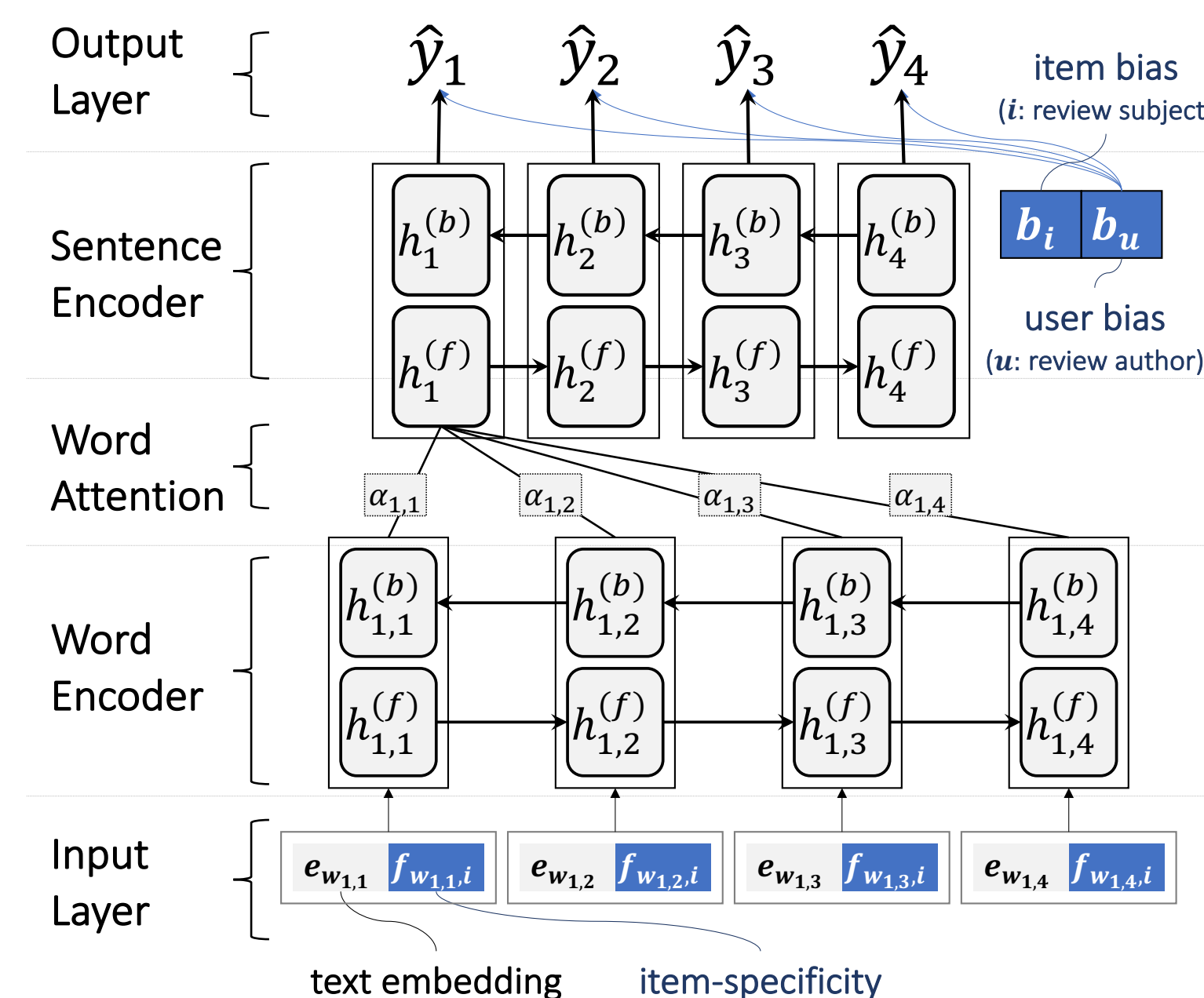


Figure: Model architecture of **SpoilerNet**

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).

	<i>Goodreads</i>		<i>TV Tropes</i>	
	AUC	AUC(d.)	AUC	Acc.
SVM	0.744	0.790	0.730	0.657
+ item-spec.	0.746 ↑	0.800 ↑	0.747 ↑	0.653 ↓
+ bias	0.864 ↑	0.793 ↑	0.722 ↓	0.536 ↓
SVM-BOW	0.692	0.729	0.756	0.702
+ item-spec.	0.693 ↑	0.734 ↑	0.774 ↑	0.710 ↑
+ bias	0.838 ↑	0.742 ↑	0.753 ↓	0.704 ↑
CNN	0.777	0.825	0.774	0.709
+ item-spec.	0.783 ↑	0.827 ↑	0.790 ↑	0.723 ↑
+ bias	0.812 ↑	0.822 ↓	0.781 ↑	0.711 ↑
- word attn.	0.898 ↓	0.880 ↓	0.760 ↓	0.695 ↓
- word init.	0.900 ↓	0.880 ↓	0.702 ↓	0.652 ↓
- sent. encoder	0.790 ↓	0.836 ↓	-	-
HAN	0.901	0.884	0.783	0.720
+ item-spec.	0.906 ↑	0.889 ↑	0.803 ↑	0.733 ↑
+ bias	0.916 ↑	0.887 ↑	0.789 ↑	0.729 ↑
SpoilerNet	0.919	0.889	0.803	0.737

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*.

Error Analysis

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

Prob.	Label	Review Text
0.35	False	Language: Low (one/two usages of d*mn)
0.32	False	Religion: None
0.39	False	Romance: None
0.59	False	Violence: Low (It's a murder mystery! Someone is killed, but it is only ever talked about.)

Table: **Distracted by Surrounding Sentences:** Review example from *The Fault in Our Stars*.

Prob.	Label	Review Text
0.08	False	This is not your typical teenage love story.
0.86	True	In fact it doesn't even have a happy ending.
0.70	False	I have to say Hazel with all her pragmatism and intelligence has won me over.
0.43	False	She is on the exact opposite side of the spectrum than characters like the hideous Bella Swan.

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

Prob.	Label	Review Text
0.01	False	The writing is simplistic, a little more so than befits even the 1st-person narrative of a 16-year-old.
0.50	True	One of things I liked best about this is having a heroine who in addition to acting for the cameras, also has to fake her affection to someone who reciprocates far more than she feels.
0.15	True	I found it very relatable.

Resources

More technical details and references can be found in our paper. Data can be downloaded from the QR code; Or from <https://github.com/MengtingWan/goodreads> Paper, data and code can also be accessed on the first author's webpage.

