Publications

Research Papers | Datasets

Research Papers

- **WSDM'20**
  - **Addressing Marketing Bias in Product Recommendations**

    Mengting Wan, Jianmo Ni, **Rishabh Misra**, Julian McAuley, in Proceedings of 2020 ACM Conference on Web Search and Data Mining (WSDM'20), Houston, TX, USA, Feb. 2020. (**15% acceptance rate**)

  - Paper | Data and Code
- **ACL'19**
  - **Fine-Grained Spoiler Detection from Large-Scale Review Corpora**

    Mengting Wan, **Rishabh Misra**, Ndapa Nakashole, Julian McAuley, in Proceedings of 57th Annual Meeting of the Association for Computational Linguistics 2019 (ACL'19), Florence, Italy, Jul. 2019. (**18% acceptance rate**)

  - Paper | Dataset | Poster | **Media**: TechCrunch, NBC, Gizmodo, Geek.com, UCSD News/UC News, TechXplore
- **RecSys'18**
  - **Decomposing Fit Semantics for Product Size Recommendation in Metric Spaces**

    Rishabh Misra, Mengting Wan, Julian McAuley, in Proceedings of 2018 ACM Conference on Recommender Systems (RecSys'18), Vancouver, Canada, Oct. 2018. (**25% acceptance rate**)

  - Paper | Code | Datasets
- **MUSE'15**
  - **Scalable Bayesian Matrix Factorization**

    Avijit Saha*, **Rishabh Misra***, Balaraman Ravindran, In Proceedings of the 6th International Workshop on Mining Ubiquitous and Social Environments (MUSE) @ PKDD/ECML, 2015 Sep 7 (pp. 43-54), Porto, Portugal. (* equal contribution)

  - Paper | Code
- **Pre-print**
  - **Scalable Variational Bayesian Factorization Machine**

    Avijit Saha, **Rishabh Misra**, Ayan Acharya, and Balaraman Ravindran.

  - Paper | Code

Datasets

- **IMDB Spoiler Dataset** [**Released**: May 2019]
  - User-generated reviews are often our first point of contact when we consider watching a movie or a TV show. However, beyond telling us the qualitative aspects about the item we want to consume, reviews may inevitably contain undesired revelatory information (i.e. 'spoilers') such as the surprising fate of a character in a movie, or identity of a murderer in a crime-suspense movie etc. For users who are interested in consuming the item but are unaware of the critical plot twists, spoilers may decrease the excitement regarding the pleasurable uncertainty and curiosity of media consumption. Therefore, a natural question is how to identify these spoilers in entertainment reviews, so that users can more effectively navigate review platforms. This dataset is collected from IMDB and contains meta-data about items as well as user reviews with information regarding whether a review contains a spoiler or not. (**500+ downloads on Kaggle**)
- **Clothing Fit Dataset for Size Recommendation** [**Released**: August 2018]
  - Product size recommendation and fit prediction are critical in order to improve customers' shopping experiences and to reduce product return rates. However, modeling customers' fit feedback is challenging due to its subtle semantics, arising from the subjective evaluation of products and imbalanced label distribution (most of the feedbacks are "Fit"). These datasets, which are the only fit related datasets available publically at this time, collected from *ModCloth* and *RentTheRunWay* could be used to address these challenges to improve the recommendation process. (**2000+ downloads on Kaggle**)
  - Please cite the following if you use the data using this link:

    ```
    Decomposing fit semantics for product size recommendation in metric spaces
    Rishabh Misra, Mengting Wan, Julian McAuley
    RecSys, 2018
    ```
- **News Headlines Dataset For Sarcasm Detection** [**Released**: June 2018]
  - Past studies in Sarcasm Detection mostly make use of Twitter datasets collected using hashtag based supervision but such datasets are noisy in terms of labels and language. Furthermore, many tweets are replies to other tweets and detecting sarcasm in these requires the availability of contextual tweets. To overcome the limitations related to noise in Twitter datasets, this **News Headlines dataset for Sarcasm Detection** is collected from two news website. *TheOnion* aims at producing sarcastic versions of current events and we collected all the headlines from News in Brief and News in Photos categories (which are sarcastic). We collect real (and non-sarcastic) news headlines from *HuffPost*. (**9000+ downloads on Kaggle**)
  - Please cite the following if you use the data using this link:

    ```
    Sarcasm Detection using Hybrid Neural Network
    Rishabh Misra, Prahal Arora
    Arxiv, August 2019
    ```
- **News Category Dataset** [**Released**: June 2018]
  - This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from *HuffPost*. This dataset could be used to produce some interesting liguistic insights about the type of language used in different news articles or to simply identify tags for untracked news articles. (**9000+ downloads on Kaggle**)

- Please cite the following if you use the data using this link:

  ```
  News Category Dataset
  Rishabh Misra
  ResearchGate, May 2018
  ```