What is on People's Minds?

Alexei Chen, Zedong Chen, Tian Xia, Harry Yang, Qisi Yang
{alexeich, zedong, tianxia, harryjy, yangqs}@umich.edu

Introduction

- Selling books is a potentially financially risky venture that requires an understanding of what topics readers are interested in
- Trends regularly change and may not be completely reflected in a single source, since different sources specialize in different topics[1]
- Our goal is to use both book reviews and social media to understand what topics matter to people and how these trends change over time, which will help us recommend emerging trends to sellers to optimize their inventory

Data Collection

- initial existing sources: us.trend-calendar.com, csv files from github.
- Scraping descriptions of books and trends from Goodreads/Google and Wikipedia, respectively, by using Selenium, Google Books API, and MediaWiki API.
- Unify data into JSON format for easier read and process
- Remove duplicates from data source
- Adopt the tf-idf to improve the algorithm

Twitter Trends on 1st January, 2022

Twitter's top trending topics on 1st January, 2022.

1. Happy New Year
2. Happy 2022
3. #HarryPotter20thAnniversary
4. Starting 2022
5. Welcome to 2022

Google Books APIs

goodreads

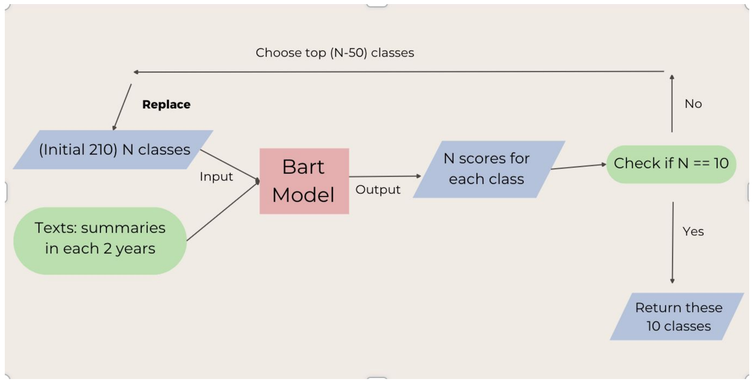
MediaWiki

Label Extraction

- We manually remove some unpromising categories such as "many" or "name"
- We utilize spaCy with the en_core_web_sm model to perform part-of-speech tagging, giving more weights on nouns
- Combining Source from all the three parts to extract 210 initial labels

Data Processing

- Start with 210 labels collected from previous part.
- We use the texts from book descriptions in each two years as input, and utilize zero-shot classification BERT to classify these texts with those labels.
- Rule out 50 labels with lowest probability each round until 10 final labels left as the algorithm showed below
- Those 10 final labels with highest probability would be considered as the 10 most popular categories.



Input	Source	Texts	Classes
	Google Books API, Goodreads, Twitter and Wikipedia	N book Google summaries	210 most important words from Goodreads summaries, Google summaries and Wikipedia description(applying techniques such as case folding, stop-word elimination, and text normalization)

Table 1: Input for model

Trend Prediction

- Use expert prompting by including "you're an expert in trend prediction" in the prompt
- Use in-context learning by providing examples of previous years' trends in the prompt
- RoBERTa used for prediction, which was trained with masked language modeling[2]

Evaluation and Results

- The evaluation results reveal that the categories deduced by the bart-large-mnli model from various book summaries often did not align closely with those determined manually.
- The second evaluation comparing trend predictions from the RoBERTa model and actual category data calculated by the BART
- Results: Based on the prediction label results, users will be able to receive the most relevant top 10 books in each category.
- Recommended books will be returned with book names, prices, author, and descriptions.

Year Range	Manual Categorization	Model Final Categorization
2017-2019	business, life, house, world, **attention**, secrets, decisions, control, power, women	classic, success, number, record, fans, championship, media, **attention**, leader, old
2015-2016	cooking, **children**, family, meals, attention, health, comedy, house, **kids**, food	advice, movement, groups, nomination, young, guide, company**children**, **kids**, program
2011-2012	adult, murder, president, campaign, history, power, death, creatures, world, **tale**	young, movement, company, position, **tale**, work, passion, action, friend, brand

Years	RoBERTa Predictions	BART Calculations
2009-2010	-	thinking, tale, host, sequel, attention, power
2011-2012	media, category, marketing, categories, news	young, power, attention, thinking, tale, drama
2013-2014	**media**, category, marketing, categories, trends	**media**, power, young, host, company, variety
2015-2016	**media**, content, category, categories, marketing	**media**, classic, power, attention, groups, success
2017-2019	**media**, content, marketing, data, media	**media**, power, areas, classic, thinking, attention
2020-2021	media, marketing, style, business, category	-

Conclusion

Future Improvements Improvements:

- Build a larger dataset and more data sources.
- Find a better book api/website. (Goodreads website is not reliable and google is lack of ratings/reviews.)
- Design a better Interface to shorten codes
- Adopt CICD to switch smoothly from one task to another.
- Merge more sophisticated models to create better predictions
- Combine user reviews & ratings from multiple sources to recommend books to users.

References

1. Althoff et al. 2013. In Proceedings of the 21st ACM international conference on Multimedia.
2. Liu et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*.