Understanding Text Classification Data and Models Using Aggregated Input Salience

Sebastian Ebert, Alice Shoshana Jakobovits, Katja Filippova {eberts,jakobovits,katjaf}@google.com

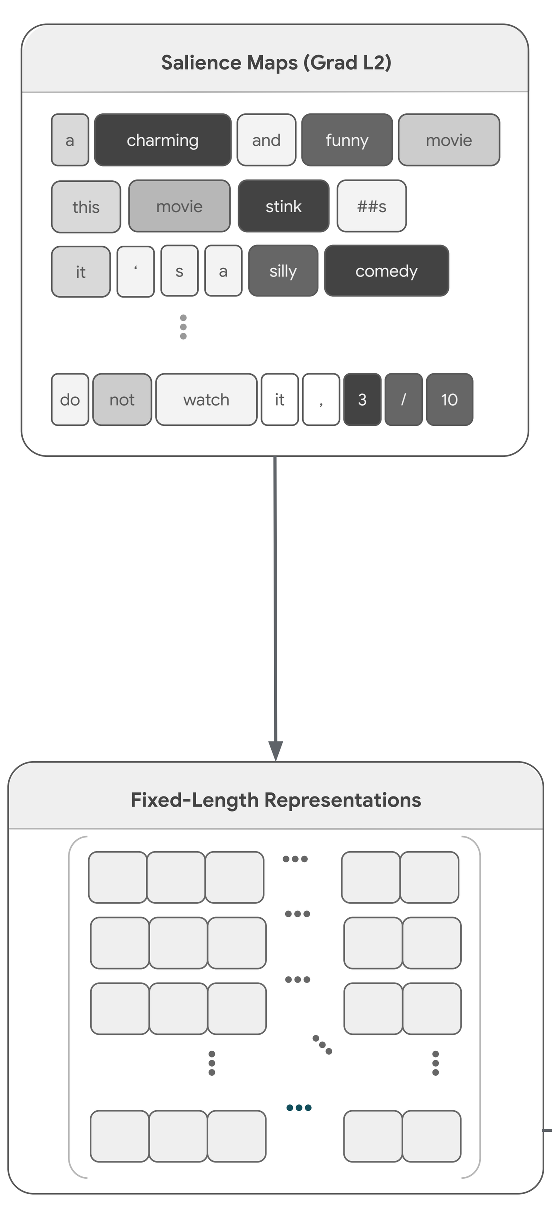
Introduction

Goal
- understand and find general patterns that a model learned from text data
- e.g., (mis-)predictions, shortcuts, etc.
- aimed at model developers

Approach
- take a method that, given a model, explains a single data example
- aggregate the explanations over an entire dataset
- aggregated explanations help gaining insights into the model and data (not just a single example)

Model Developer's Questions
- What patterns did the model learn from the data? (this poster)
- What is the model sensitive to? (paper)
- How can puzzling predictions be explained? (paper)

Method



Single Example Input Salience
- gives an importance weight to every input token
- gradient-based method, here: Grad L2 for BERT

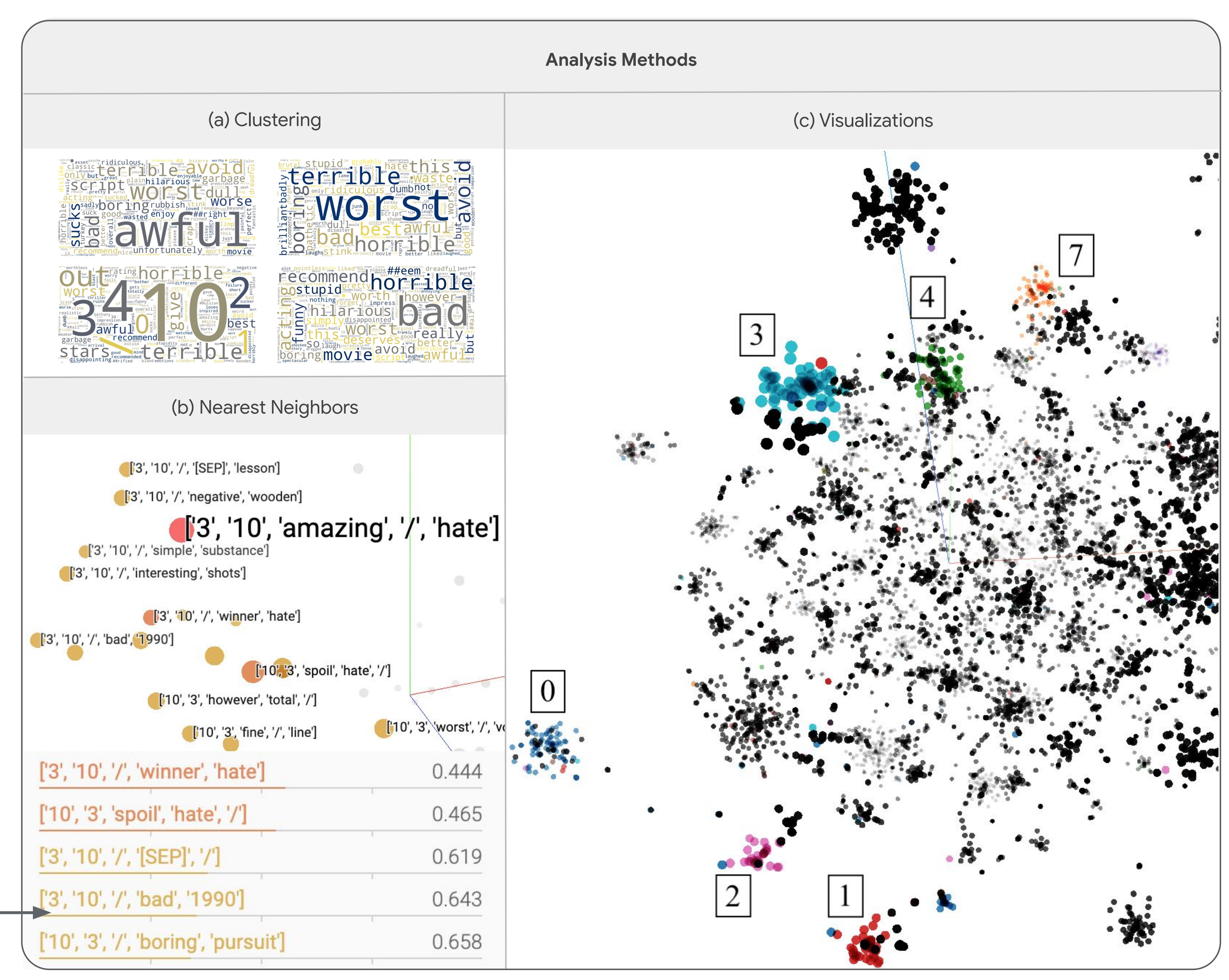
Problems of Single Example Explanations
- may lead the developer to discover false generalizations ("not" as indicator of contradiction in NLI)
- time-consuming to go through a large number of inputs to find patterns
- may be difficult to spot unintuitive patterns from individual examples even when the salience maps hints at its presence

Solution: aggregation of fixed-length representations

Baseline Representations
- vocabulary with token PMI weighting
- average BERT word piece embedding
- CLS encoding of BERT model

Salience Representations (ours)
- vocabulary with salience weighting
- sum of word piece embedding weighted by salience

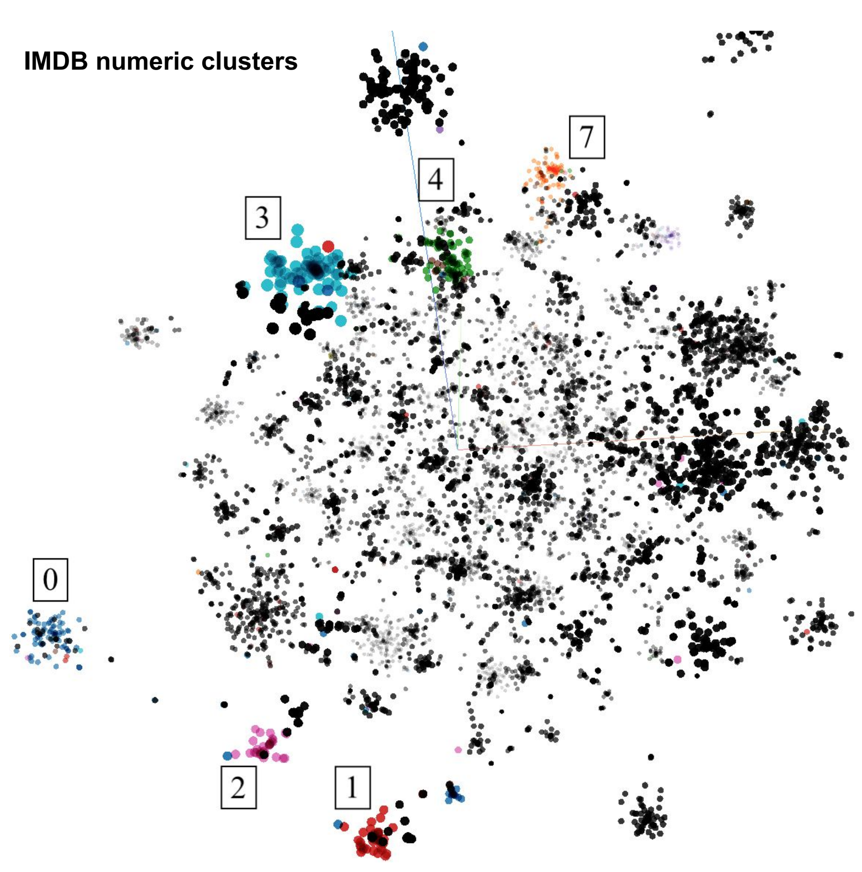
Analysis Methods: clustering, nearest neighbors, visualizations



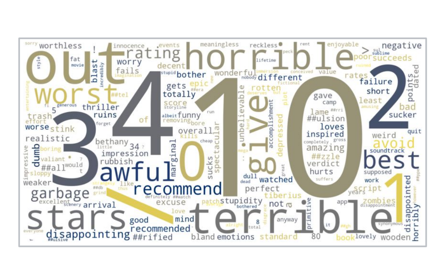
Results

Question: What patterns did the model learn from the data?
- IMDB: movie reviews
- binary polarity classification (negative, positive)

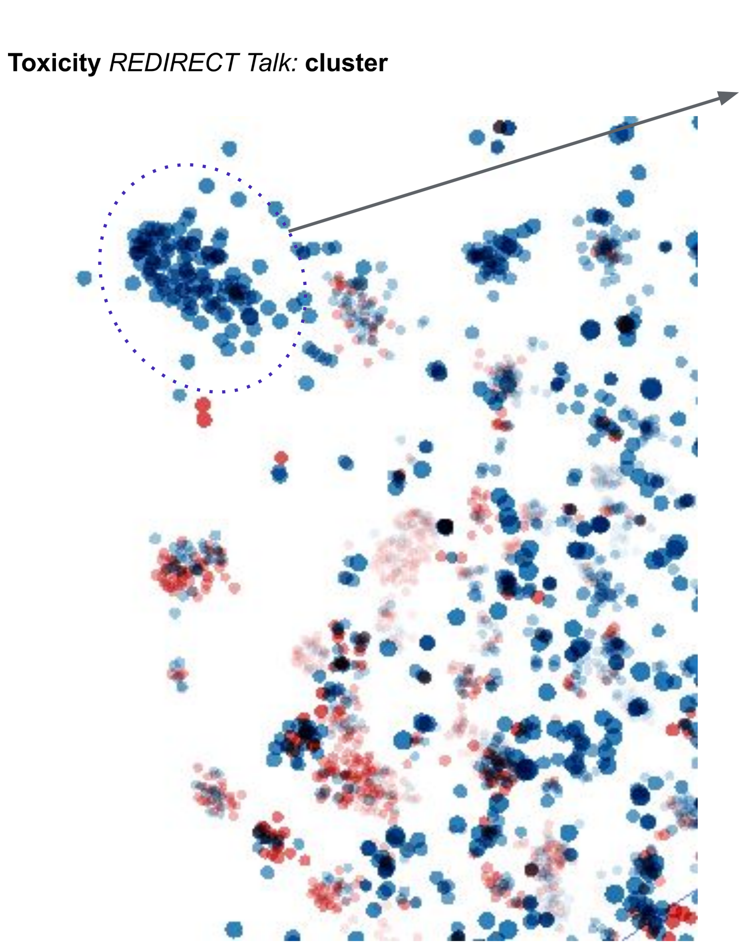
- t-SNE on fixed-length representation reveals interesting artifact
- e.g., bottom center examples (red) contain "1"
- --> many examples contain "1 / 10 stars","3 out of 10","4 / 10"
- this is a potential shortcut



- find numeric patterns with regexes (245/2500 examples)
- cluster representation matrix
- evaluate representations based on highest precision cluster (out of 5) (see word cloud of such a cluster)
- prec: ratio of numeric examples within the cluster
- recall: ratio of numeric examples in the cluster out of all numeric examples



Representation	Size	Prec.	Recall
B1: PMI vocab	.16	.12	.19
B2: avg emb	.08	.16	.13
B3: CLS-encoding	.47	.12	**.58**
S1: salience vocab	.05	**.98**	.50
S2: salience emb	.05	.93	.53



redirect talk : 2011 ffas senior league
redirect talk : hurricane ismael (1983)
redirect talk : we ' ll be together
redirect talk : the pier shops at caesars
redirect talk : bill kern (tackle)
redirect talk : green to gold (book)

- Toxicity classification
- revealed pattern: "REDIRECT Talk: {Article Title}"
- contentless comments, contain no information about toxicity
- 1-1.5% of Toxicity data follows the same pattern and is useless

Conclusion

- We propose **aggregated salience** representations for discovering patterns that a model learned from the data.
- We can answer distinct, yet **common model developer questions**.
  - What patterns did the model learn from the data?
  - What is the model sensitive to?
  - How can puzzling predictions be explained?
- The answers are a first step for improving the model.
- Use clustering and visualization on **your own models** with LIT and the embedding projector.