MSiA 414: Natural Language Processing Independent Project

Literature review

Paper 1:

Rania Albalawi, Tet H. Yeap and Morad Benyoucef. 2020. *Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis*. https://doi.org/10.3389/frai.2020.00042

Evaluates different methods for topic modeling:

1. Latent semantic analysis (LSA):
    a. Pros: reduces dimensionality of TF-IDF by using single value decomposition
    b. Cons: requires human judgement in setting number of topics, does not capture correlation across topics
2. Latent Dirichlet allocation (LDA):
    a. Pros: Requires no training data, handles mixed-length documents
    b. Cons: Unable to model relations among topics that help understand deep structure
3. Non-negative matrix factorization (NMF):
    a. Pros: Fast process (for real time data), can extract meaningful topics with no prior knowledge of underlying meaning
    b. Cons: Risk of semantically incorrect results
4. Random projection (RP)
    a. Pros: Robust in imbalanced datasets and high-dimensionality data
    b. Cons: Slow, sensitive to noise and applicable to only subset of data
5. Principal component analysis (PCA)
    a. Pros: Output can be easily visualized, low noise sensitivity
    b. Cons: Covariance matrix is difficult to evaluate, expensive to compute

Two textual datasets are used to compare these approaches, using traditional statistical evaluation methods: recall, precision F1-score. This paper concludes that LDA and NMF produced the highest quality results (along with the most valuable outputs from a diversity and meaning perspective).

Paper 2:

Marco T. Ribeiro, Sameer Singh and Carlos Guestrin. 2016. *"Why Should I Trust You?" Explaining the Predictions of Any Classifier.* https://arxiv.org/pdf/1602.04938.pdf

This paper proposes the LIME explanation technique (used to increase trust in black box models). The LIME framework can explain the predictions from any classifier.

LIME works by choosing a single prediction to be explained, then creates permutations in the data for this given instance and collects results. It then weighs the new samples based on how closely they match the original data prediction. Finally, a less complex (and interpretable) model is used to explain this local instance.

Not Only can LIME be used to "lift the hood" over a given observation, highlighting why it was classified to one category vs the other; the comparison of different LIME outputs can be used as a new tool to select between models (in addition to traditional statistical methods).

Paper 3:

Molnar, Christoph. 2019. *Interpretable machine learning. A Guide for Making Black Box Models Explainable,* Chapter 9: Local Model-Agnostic Methods. https://christophm.github.io/interpretable-ml-book/

Advantages of LIME:

- LIME is one of the only methods which works for tabular, text and image data
- LIME's ease of use is second to none
- It produces a 'fidelity' measure which evaluates how well the local model approximates the black box prediction

Disadvantages of LIME:

- The sampling of data points for one observation does not take into account the correlation between data points. A local explanation model may therefore be biased towards learning
- Instability of explanations: repeating the sampling process may result in different explanations

Paper 4:

Ian Tenney, James Wexler et al. 2020. *The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models.* https://arxiv.org/pdf/1602.04938.pdf

This paper presents the Language Interpretability Tool (LIT), a toolkit and browser-based interface (UI) for NLP model understanding. LIT supports both local explanations (similar to LIME), but also aggregate analysis (metrics, embedding spaces, etc).

In the interface, users can:

- Explore the dataset (by rotating projections, zooming in on clusters, even identifying clusters which can be further investigated)
- Users can identify, save and then analyze interesting data points or clusters of data points (outlying clusters in the embedding space)
- Explain local behavior (LIME type exploration of predictions for one data point)
- Generate new data points (see if change of one given word affects classification)
- Compare models side by side

Some interesting applications of LIT include:

- Sentiment analysis: how does sentiment change when changing one word in a text entry (using LIT's ability to modify or create new datapoints)
- Detection of Bias: does classification change if male references are changed to female, et

Paper 5:

Xiaowei Zhang, Shuchen Qiao, Yang Yang and Ziqiong Zhang. 2020. *Exploring the Impact of Personalized Management Responses on Tourists' Satisfaction: A Topic Matching Perspective*.

https://doi.org/10.1016/j.tourman.2019.103953

With the increasing frequency of online reviews, managers in the tourism industry are slowly moving from responding with general apologies to poor replies to more customized "context specific" responses. To this day, various different techniques have been tested to respond to negative feedback (pology, redress, facilitation,

credibility, problem-solving, courtesy, explanations), however the actual topic of the complaints has often been overlooked. Some have suggested that management responses without acknowledgement of context may be detrimental.

This paper reviews current literature, highlighting a few interesting points:

- The value of eWOM (electronic word of mouth), typically evaluated as mean rating and variance of rating. eWOM has been demonstrated to have an impact on hotel response styles to online comments, and even on hotel financial performance
- Topic matched responses may heighten customers' perceived effectiveness of hotel responses

The paper goes on to identify a subset of general topics used in hotel evaluation (food & beverage, price, facilities, service, environment) and evaluates their impact on eWOM.

The approach to selecting topics was:

1. Identification of high-frequency words (excluding stop-words)
2. Clustering was used among the high frequency words into a subset of clusters

An SVM model was then used to classify/assign each text entry to a topic that was chosen. The main variable computed in this study was then the average degree of **matching** between the **customer complaint topic** and the **hotel response topic** in modeling hotel performance.