ISEN 685 Final Report: Disentangling content and opinion using structured variational topic model

Ximin Yue, Ran Wei

Introduction

Recent years have seen a surge in the development and deployment of automated vehicle (AV) technologies, ranging from semi-autonomous vehicles to advanced automated driver assistance systems. In order to facilitate public acceptance and appropriate use of AV technologies, it is importance to align public expectation of AV technologies with their current state of capabilities. Drivers' misunderstandings of AV behavior have led to a number of fatal crashes that could otherwise have been prevented. An analysis of the California police report on AV crashes shows that the majority of crashes are due to the lack of transparency of the systems' behavior to either the driver or the surrounding vehicles. One way to bilaterally improve AV system transparency and calibrate public expectation is through understanding the current state of public (mis)conception of AV technologies, which can be used to device strategies to (mitigate)promote these (mis)conceptions.

Social media sites serve as idea platforms to assess public perception and expectation of AV technologies. With the rise of social media platforms brought by internet technology, people are more likely to use social media as a channel to express their real perspective and ideals, as the internet has the characteristics of virtual, penetration, divergence, concealment and randomness. Therefore, the project aims to model the public opinions of AV on twitter which is one of the most influential social media platforms of the world. The result of this data mining research can help us to enhance the understanding of the public sentiment on AV and offer guidance for the related company and department to improve AV safety.

Topic modeling. Topic modeling is the most popular approach to categorize large corpus of text data into a small number of topics, each representing a meaningful area of discussion. Using Latent Dirichlet Allocation (LDA), a probabilistic topic modeling approach, Wei et al. identified 6 topics from Twitter discussions after AV crashes and technology advancement events. These topics range from AV safety and attribution of fault after a crash event to market and sales and urban mobility.

While LDA is able to summarize tweets according to their content, it fails to uncover more nuanced information such as the users' stance and opinions, which are important for understanding their perceptions and expectations. In addition, LDA suffers from learning from short text documents such as tweets and large vocabularies, as these two properties combined lead to high sparsity in the dataset. To uncover opinionated topics, we use tweet-level sentiment scores as additional data to guide the learning and inference of the topic model as in (Card et al., 2017; Pergola et al., 2020). To overcome data sparsity, we represent tweet documents with word-embeddings, which project discrete word tokens into a condensed semantic space (Dieng et al., 2020).

Word embedding. Word embedding is an alternative way of representing words as fixed-length vectors so that similar words are close-by in the vector space. This representation has the advantage

of reducing the number of parameters and maintaining semantic similarity of words compared to the bag of words representation. Word embeddings are usually trained by predicting a word given its surrounding words, i.e., the context, using log-bilinear models. When trained on large generic corpus, such as Wikipedia, the trained word embeddings can be used on a different corpus and perform a different set of tasks, such as text classification. Domain and task-specific word embedding can be trained by using domain specific corpus and task-specific information, such as sentiment labels (Maas et al., 2011).

A number of works have suggested the close connection between topic modeling and word embedding learning (Dieng et al., 2020; Foulds, 2018; Maas et al., 2011). Specifically, the latent topic in the LDA model can be considered as the context the word embedding models, and the surrounding words in a word embedding model can be considered the topic for that word. Maas et al. (2011) used this insight to train word embeddings using a modified topic model where topics are represented as continuous instead of discrete distributions. In contrast, Foulds (2018) proposed a method to train word embedding for small corpus by first training a topic model and then use the learned topics as the context in word embedding models. Dieng et al. (2020) suggested that topic and word embedding can be trained together in the variational inference algorithm, where they used a simple MLP model as the variational distribution. They found that pretrained word embeddings outperform word embeddings trained with the topic model. However, pretrained word embedding has the disadvantage of being not domain specific. We experimented with the MLP inference model proposed by Dieng et al. (2020) with both pretrained word emebddings and word embeddings trained from scratch and found that the topic model is highly prone to mode collapse where all learned topics are the same. We hypothesize this is due to a lack of learning signal for the word embeddings and show that using embeddings in the inference model greatly improves inference and embedding learning performance.

Disentangled representations. Disentangled representations refer to factors that are independent of each other in a generative process. Topics are usually products of entangled factors, such as contents and opinions, however, these entangled factors are usually not discovered by topic models without modification of model structures. Earlier works such as supervised LDA (Blei and McAuliffe, 2010) have used sentiment labels to extract topics imbued with sentiment. However, in applications such as product review analysis, it is useful to separate topics from sentiment in order to understand the cause of positive and negative reviews of the products (Esmaeili et al., 2019; Pergola et al., 2020). A number of works (John et al., 2018; Pergola et al., 2020) use adversarial methods to disentangle sentiment from content, however, these methods deviate from principled probabilistic generative models. In contrast, we disentangle sentiment from content by designing a structured model to control the representations being learned.

Structured variational topic model

Generative process. The goal of the structured topic model is to disentangle opinion from content in text corpus. We assume the text corpus is generated from a fixed number of topics K. The author's opinion towards the topics is modeled by a single continuous variable s. We assume access to document level sentiment labels $y \in [-1, 1]$ which gives us indirect information about the authors' opinions. We model the text corpus and sentiment labels with the following generative process:

1. Independently draw topic proportion $\theta \sim P(\theta)$ and opinion $s \sim P(s)$
2. For each word in a document, independently draw topic $t \sim P(t|\theta)$ and then word $w \sim P(w|t, s)$

3. Draw document sentiment label from $y \sim P(y|\theta, s)$

where the topic proportion θ is a continuous vector restricted to a probability simplex (same as LDA).

The log-marginal probability of a document with N words is:

$$\begin{aligned} \log P(w_{1:N}, y) &= \log \int P(s) \int P(\theta) P(y|\theta, s) \prod_{t=1}^{N} P(w_t|\theta, s) d\theta ds \\ &= \log \int P(s) \int P(\theta) P(y|\theta, s) \prod_{i=1}^{N} \sum_{t} P(t|\theta) P(w_i|t, s) d\theta ds \end{aligned} \tag{1}$$

We model the prior over topic proportion as a standard logistic-normal distribution and the prior over sentiment as a standard normal distribution. Since the integration over the continuous variable s is intractable, we follow (Srivastava and Sutton, 2017) and maximize a lower bound of (1) using a Auto-Encoding Variational Bayes (AEVB) algorithm. Parameterizing a variational distribution $Q(\theta, s|w_{1:N})$ from the same family of the priors, the log marginal likelihood lower bound is well known:

$$\mathcal{L} = \mathbb{E}_{Q(\theta,s|w_{1:N})}[\log P(y|\theta, s) + \sum_{i=1}^{N} \log P(w_i|\theta, s)] - D_{KL}[Q(\theta, s)||P(\theta, s)] \tag{2}$$

where the expectations can be estimated from samples from the variational distribution.

Structured model. The emphasis of the structured topic model is the parameterization of the conditional distributions $P(w|t, s)$ and $P(y|\theta, s)$. We desire the parameterization to be interpretable so that they need not be analyzed with post-hoc methods. We also desire the latent variables to be semantically meaningful and consistent across topics, such that by varying s from negative to positive, we can visualize the word conditional distributions across different topics.

We propose to use the following parameterization for the word conditional distribution:

$$\begin{aligned} P(w_i|t_k, s) &= \frac{\exp(f(w_i, t_k, s))}{\sum_j \exp(f(w_j, t_k, s))} \\ f(w_j, t_k, s) &= b_j + \overrightarrow{w_j}^T \overrightarrow{t_k} + s\overrightarrow{w_j}^T \overrightarrow{h_k} \end{aligned} \tag{3}$$

where w_j is the embedding for word j, t_k is the embedding for topic k, h_k is the sentiment embedding for topic k, and b_j is the background log-frequency for word j. Thus higher word frequencies are modeled by higher values of inner products between the word embedding and topic and sentiment embeddings. The first two terms model the nominal word probabilities when the opinions for the topics is neutral, i.e. $s = 0$. However, when s is positive or negative, words that have higher inner product values with the sentiment embedding are assigned higher frequencies. In addition, we use a separate sentiment embedding for every topic rather than a shared sentiment embedding for all topics. This is because we are interested in extracting words associated with the specific positive and negative opinions for each topic rather than the extracting words that have high positive or negative sentiment universally.

Since the sentiment labels are bounded continuous values, we model the distribution over sentiment as beta distributions $Beta(\alpha, \beta)$ with parameters α and β. Higher values α/β leads to higher/lower mean sentiment, and higher or lower values of both parameters lead to higher precision of the sentiment distribution. We thus use a similar parameterization as the word conditional distribution:

$$\begin{aligned} \alpha &= \exp(x_\alpha + s y_\alpha^2) \\ \beta &= \exp(x_\beta - s y_\beta^2) \end{aligned} \tag{4}$$

where $x_\alpha, x_\beta \in \mathbb{R}^K$ are the baseline log-value of the α and β parameters. $y_\alpha, y_\beta in \mathbb{R}^K$ are the sentiment weights of the α and β parameters. We square the sentiment weights so that positive values of s can only increase α and decrease β, and vise-versa for negative values of s.

Experiment

The experiment includes data pre-processing, implementing LDA, word vectors and calculation sentiment scores of each topic, implementing LDA, BOW and calculation sentiment scores of each topic and implementing ETM, BOW and calculation sentiment scores of each topic and implementing ETM, word to vectors and calculation sentiment scores of each topic. Finally, we tried to introduce a parameter "query" which ranges from positive infinity to negative infinity. Where a query is a parameter which can increase(when it was positive) or decrease(when it was negative) the right of positive or negative topics relatively so that we can check the different sentiment and opinion of each topic. The method is called disentangled ETM.

1 Data and pre-processing

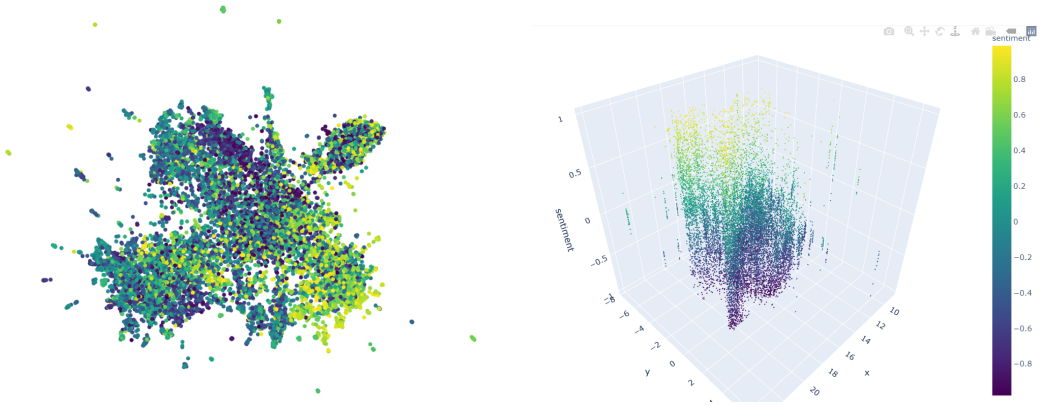
The dataset was collected in February, 2020 from Twitter's website. It was captured for 12 AV events from 2014 to 2019. For each event a separate CLL search was conducted for the 10 days before and after each crash event.

The raw data were pre-processed through removing hyperlinks, tags and hashtags, pictures and emojis. Then, we extracted the word tokens and retained the nouns, verbs and adjectives and only kept their morphological roots. Then, the stop words (words of common high frequency but low contextual meaning, "a", "the", "is", "are" and etc.) Finally, we removed the duplicate tweets and the tweets that contained less than 3 words. Therefore, we get the pre-processed data.

Next step, we used LDA and BOW to extract 10 topics of the pre-processed data. There are 2 of the topics relevant to AV, we extract their tweets relatively. Finally, we use logistic regression to distinguish between the AV tweets and the non-AV tweets of the tweets from these two topics. To summarize and map the data, we use Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) to visualize the processed AV data.

Every single point represents every processed AV tweet. We can easily look up the content and sentiment of each tweet by this data visualization process. From the top view, (Figure 1(a)) the tweets with similar topics cluster together. The color shows the sentiment of each tweet. The tweets are more positive and have brighter colors. The Z axis shows the sentiment of the tweets. From the 3D view (Figure 1(b)), we can check the different opinions and sentiment of people on the same AV topic. We preliminarily analyzed the composition and distribution of the data.

We also use K-means to preliminarily cluster 4 classes of different topics that are shown in different



(a) UMAP 2D (b) UMAP 3D

Figure 1: UMAP for AV tweets

colors.(Figure 2) We found that the top left of the map (Figure 1(a)) is relative to the tweets of the Uber. Bottom left and bottom right are discussions about Tesla and other companies related to AV respectively. Therefore, we got a baseline to verifying the results and the effectiveness of our model.

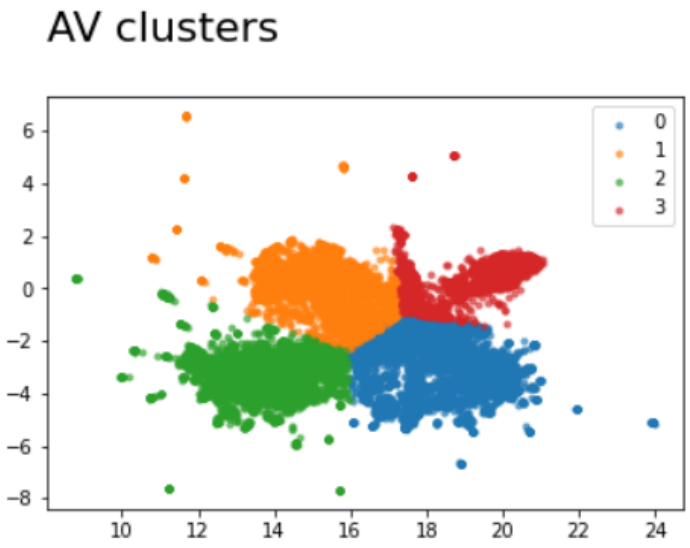


Figure 2: cluster: AV tweets

2 Experiment procedure

We did the experiment on several different models and parameters combinations. The grid search was used to find the best parameter combinations including "sample ratio", "train ratio", "batch size", "min df", "max vocab", "num topics" and "hidden dim". The model includes LDA with BOW, LDA with word vectors and ETM with embedding. Then, we compared the performance of each model. The criterion includes the test error, the topic diversity(td), the topic coherence(tc) and the evidence lower bound (ELBO).

Results and conclusions

After comparison, from the result (Figure 3) the best model is not using proLDA and training the etm for its lowest ELBO and relatively fine tc and td.

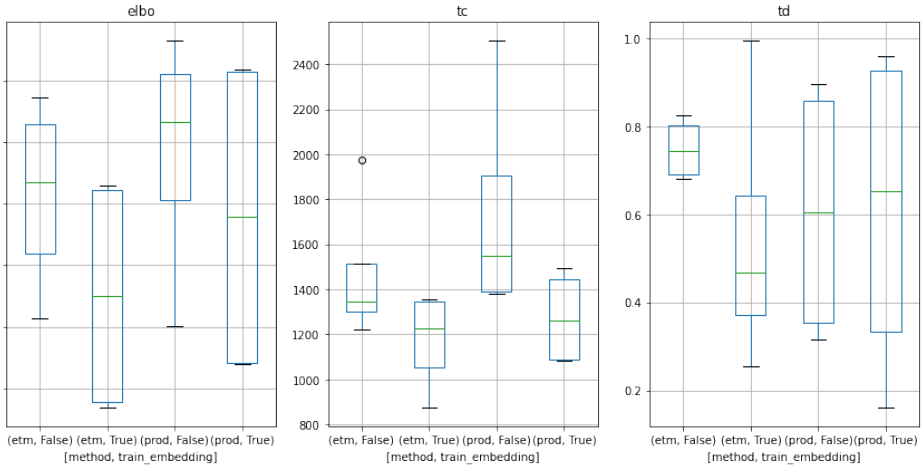


Figure 3: Result of best model selection

From the result, (Figure 4) when the topic number is 4 the model has the best performance for its lowest ELBO and relatively fine tc and td. Also, it has the best sentiment diversity.

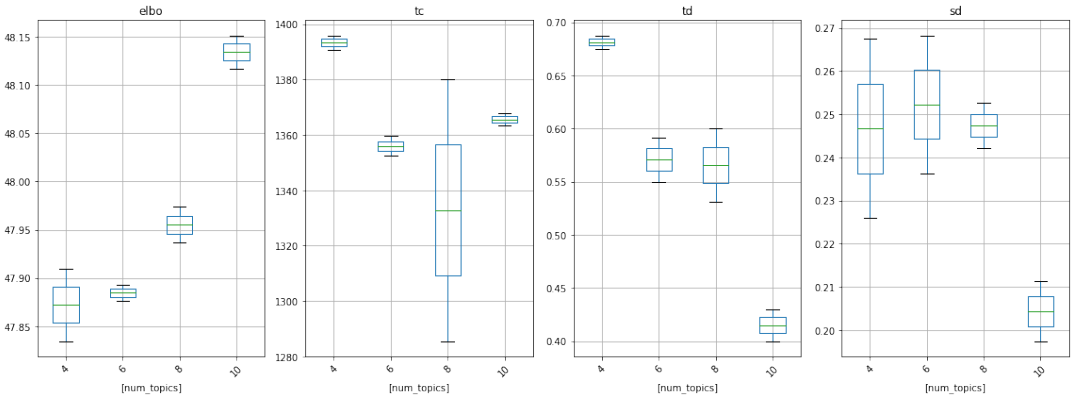


Figure 4: Result of best model selection

From the figure 5. We can conclude these 4 main topics from the dataset, which are exactly same as the results we learned from the UMAP data visualization.

The topic term and topic sentiment are showing below. (Figure 6) When we set the query to -1, the sentiment for each topic decreased. It shows more negative words of the topic. However, When we set the query to 1, the topic sentiment for each topic increase. It shows more positive words of the topic. (Figure 7)

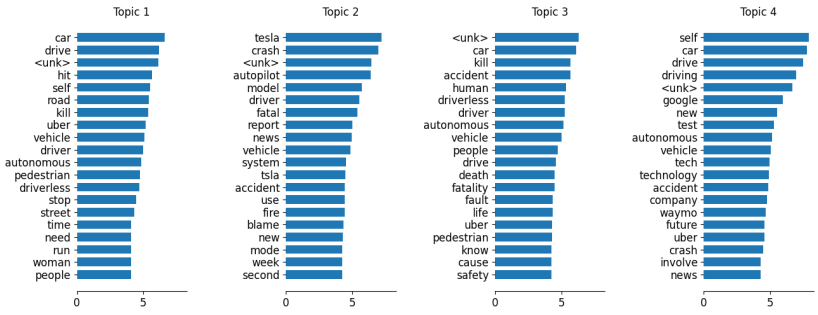


Figure 5: Result of topic number selection

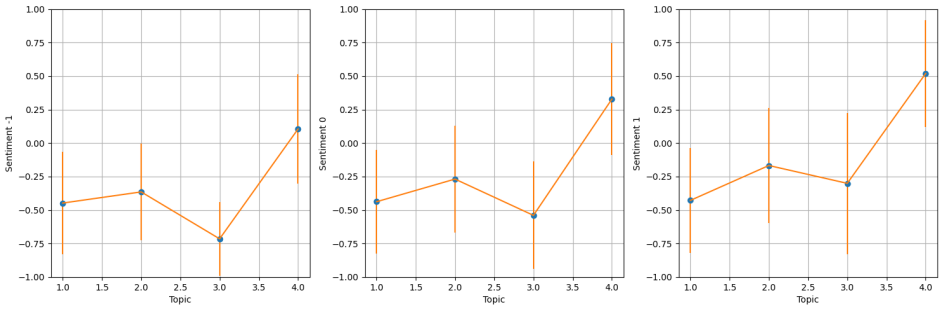
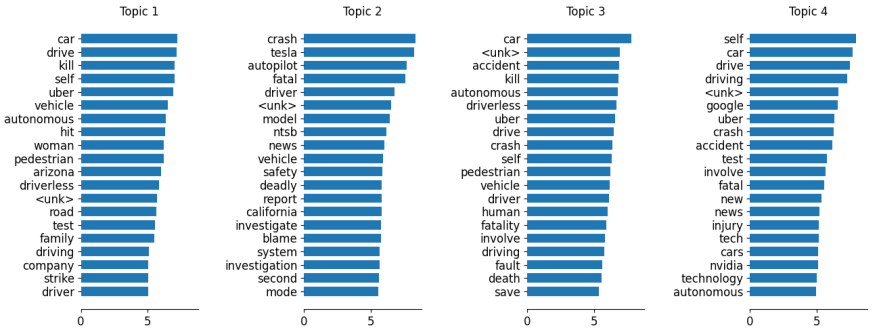
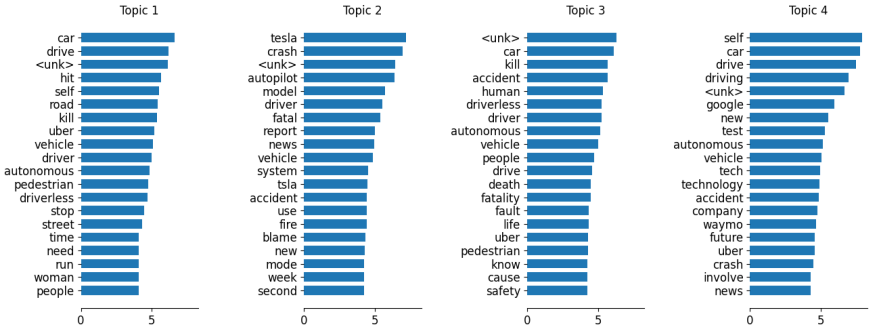


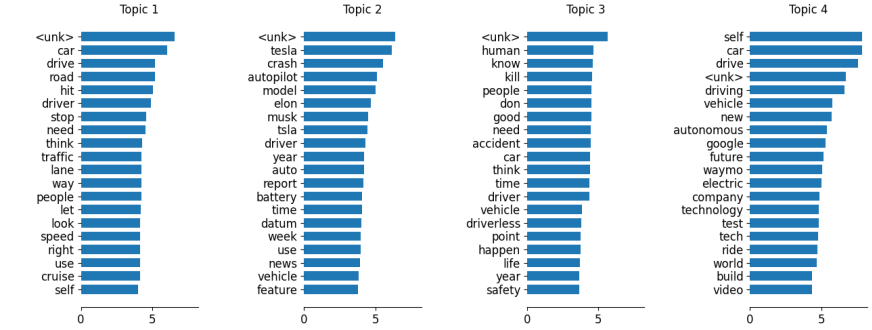
Figure 6: Change of the sentiment with different queries



(a) Query = -1



(b) Query = 0



(c) Query = 1

Figure 7: Topics with different query

References

Blei, D. M., & McAuliffe, J. D. (2010). Supervised topic models. *arXiv preprint arXiv:1003.0783*.

Card, D., Tan, C., & Smith, N. A. (2017). Neural models for documents with metadata. *arXiv preprint arXiv:1705.09296*.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453.

Esmaeili, B., Huang, H., Wallace, B., & van de Meent, J.-W. (2019). Structured neural topic models for reviews. *The 22nd International Conference on Artificial Intelligence and Statistics*, 3429–3439.

Foulds, J. (2018). Mixed membership word embeddings for computational social science. *International Conference on Artificial Intelligence and Statistics*, 86–95.

John, V., Mou, L., Bahuleyan, H., & Vechtomova, O. (2018). Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.

Pergola, G., Gui, L., & He, Y. (2020). A disentangled adversarial neural topic model for separating opinions from plots in user reviews. *arXiv preprint arXiv:2010.11384*.

Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.