Abstract

Mining public opinions about vaccines from social media has been increasingly relevant to analyse trends in public debates and to provide quick insights to policy-makers. However, the application of existing models has been hindered by the wide variety of users’ attitudes and the new aspects continuously arising in the public debate. Existing approaches, frequently framed via well-known tasks, such as aspect classification or text span detection, make direct usage of the supervision information constraining the models to predefined aspect classes, while still not distinguishing those aspects from users’ stances. As a result, this has significantly hindered the dynamic integration of new aspects. We thus propose a model, namely Disentangled Opinion Clustering (DOC), for vaccination opinion mining from social media. DOC is able to disentangle users’ stances from opinions via a disentangling attention mechanism and a Swapping-Autoencoder, and is designed to process unseen aspect categories via a clustering approach, leveraging clustering-friendly representations induced by out-of-the-box Sentence-BERT encodings and disentangling mechanisms. We conduct a thorough experimental assessment demonstrating the benefit of the disentangling mechanisms and cluster-based approach on both the quality of aspect clusters and the generalization across new aspect categories, outperforming existing methodologies on aspect-based opinion mining.

1 Introduction

Mining public opinions about vaccines from social media has been hindered by the wide variety of users’ attitudes and the continuously new aspects arising in the public debate of vaccination (Husain et al., 2021). The most recent approaches have adopted holistic frameworks built on morality analysis (Pacheco et al., 2022) or neural-based models predicting users’ stances on different aspects of the online debate (Zhu et al., 2022). So far, these frameworks have been frequently framed via well-known tasks, such as aspect classification or text span detection, that use supervision to train text classifiers. However, such a direct usage of the supervision information has constrained the models to predefined aspect classes and restricted their flexibility in generalising to opinions with aspects never seen before (e.g., new moral issues or immunity level).

To mitigate this limitation, some of the most promising approaches have been devised as supervised models generating clustering-friendly representations (Tao et al., 2021). These have recently shown promising results on open-domain tasks when combined with pre-trained language models (PLM) thanks to their flexibility, generalisation, and need for minimal tweaks (Reimers and Gurevych, 2019; Sircar et al., 2022). However, despite the improved capabilities in capturing the overall text semantics, existing models for text clustering (Miranda et al., 2018; Meng et al., 2019; Shen et al., 2021; Zhang et al., 2021a), still struggles to distinguish between the mixed users’ stances and aspects on vaccination, and as a result, they often generate clusters that do not reflect the novel aspects of interest. As an illustrating example, consider the tweets “mRNA vaccines are poison” and “The Pfizer vaccine is safe”, that the majority of existing methodologies are prone to cluster into different groups due to the opposite stances manifested, despite the fact that both of them are targeting safety issues.

To address the aforementioned problem, we posit that a model should be able to (i) disentangle the stance from the aspect discussed, and simultaneously (ii) use the generated representations in a framework (e.g., clustering) that ease the integration of aspects never seen before. We thus propose a novel representation learning approach, called the Disentangled Opinion Clustering (DOC) model, which performs disentan-
gled learning (Mathieu et al., 2016) via text autoencoders (Bowman et al., 2016; Montero et al., 2021), and generates clustering-friendly representations suitable for the integration of novel aspects\(^1\). The proposed model, DOC, learns clustering-friendly representations through a denoising autoencoder (Montero et al., 2021) driven by out-of-the-box Sentence-BERT embeddings (Reimers and Gurevych, 2019), and disentangles stance from opinions by using the supervision signal to drive a disentangled cross-attention mechanism and a Swapping Autoencoder (Park et al., 2020).

We conducted an experimental assessment on two publicly available datasets on vaccination opinion mining, the Covid-Moral-Foundation (CMF) (Pacheco et al., 2022) and the Vaccination Attitude Detection (VAD) corpora (Zhu et al., 2022). We first assessed the quality of the disentangled representation in generating aspect-coherent clusters. Then, we measured the generalisation of the proposed approach via a cross-dataset evaluation by performing clustering on a novel dataset with unknown aspect categories. Finally, we showed the benefit of this approach on the traditional stance classification task, along with a report on the thorough ablation study highlighting the impact of each model component on the clustering quality and the degree of disentanglement of the generated representations.

Our contributions can be summarized as follows:

- We introduce DOC, a Disentangled Opinion Clustering model to generate clustering-friendly representations, which distinguishes between users’ stances and opinions in the vaccination debate and integrates newly arising aspects via a clustering approach.

- Unlike traditional aspect-based classification models, we outline a framework adopting limited supervised signals provided by few stance and aspect labels, functioning as inductive biases to generate clustering-friendly representations.

- We conduct a thorough experimental analysis on the two major publicly available datasets on vaccination opinion mining from social media, and demonstrate the benefit of the disentangling mechanisms on the quality of aspect clusters, the generalization across datasets with different aspect categories, and the traditional stance classification task.

2 Related Work

Sentence Bottleneck Representation  Sentence representation learning typically aims to generate a fixed-sized latent vector that encodes a sentence into a low-dimensional space. In recent years, in the wake of the wide application of pre-trained language models (PLMs), several approaches have been developed leveraging the PLMs to encode sentence semantics. The most prevalent work is the SBERT (Reimers and Gurevych, 2019) that fine-tunes BERT (Devlin et al., 2019) on the SNLI dataset (Bowman et al., 2015) through a siamese pooling structure. The learned representations are immediately applicable to a wide range of tasks, such as information retrieval and clustering, significantly reducing the effort required to generate the task-specific representations (Thakur et al., 2021).

More recently, Montero et al. (2021) presented a sentence bottleneck autoencoder, called AutoBot, that learns a latent code by reconstructing the perturbed text. Their model indicates the importance of topic labels as reconstruction objectives.

Disentangled Latent Representation  Earlier works explored disentangled representation to facilitate domain adaptation (Bengio et al., 2013; Kingma et al., 2014; Mathieu et al., 2016). In recent years, John et al. (2019) generated disentangled representations geared to transfer holistic style such as tone and theme in text generation. Park et al. (2020) proposed the Swapping autoencoder to separate texture encoding from structure vectors in image editing. The input images are formed in pairs to induce the model to discern the variation (e.g., structure) while retaining the common property (e.g., texture). However, recent studies show that disentanglement in the latent space is theoretically unachievable without access to some inductive bias (Locatello et al., 2019). It is suggested that local isometry between variables of interest is sufficient to establish a connection between the observed variable and the latent variable (Locatello et al., 2020a; Horan et al., 2021), even with few annotations (Locatello et al., 2020b). This is in line with (Reimers and Gurevych, 2019; Lu et al., 2022) where contrastive pairs are leveraged for training, which illuminates our work to utilize labels and reconstruction of perturbed text to induce the disentanglement.

\(^1\)The code and model are available at https://github.com/somethingx1202/DOC.
Text Clustering The recent development in neural architectures has reshaped clustering practices (Xie et al., 2016). For example, Zhang et al. (2021b) leveraged transformer encoders for clustering over the user intents. Several methods utilised PLM embeddings to discover topics which were subsequently used for clustering news articles and product reviews (Huang et al., 2020; Meng et al., 2022). Others exploited the neural components, i.e., the BiLSTM-CNN (Zhang et al., 2019), the CNN-Attention (Goswami et al., 2020) and the Self-Attention (Zhang et al., 2021c) to offer end-to-end clustering. Zhang et al. (2021a) developed the Supporting Clustering with Contrastive Learning (SCCL) model by augmenting the disparity between short text. A notable work is DS-Clustering (Sircar et al., 2022), which extracts aspect phrases first then clusters the aspect embeddings. Outside of clustering methods, there is a surging interest in clustering-friendly representations (Tao et al., 2021). Yet, few methods cluster documents along a particular axis or provide disentangled representations to cluster over a subspace.

Vaccination Opinion Mining The task of vaccination opinion mining is commonly carried out on social media to detect user attitudes and provide insights to be used against the related ‘infodemic’ (Kunneman et al., 2020; Wang et al., 2021; Chandrasekaran et al., 2022; Zhao et al., 2023). Recent approaches rely on semantic matching and stance classification with extensions including human-in-the-loop protocols and text span prediction to scale to the growing amount of text (Pacheco et al., 2022; Zhu et al., 2022).

3 Methodology

We build our approach upon two vaccination opinion corpora (Pacheco et al., 2022; Zhu et al., 2022). In both corpora, a small number of tweets are labelled, each of which is annotated with a stance label (‘pro-vaccine’, ‘anti-vaccine’ and ‘neutral’) and a text span or an argumentative pattern denoting an aspect. For example, for the tweet, ‘The Pfizer vaccine is safe.’, its stance label is ‘pro-vaccine’ and the argumentative pattern is ‘vaccine safety’. Since vaccination opinions explode over time, supervised classifiers or aspect extractors would soon become outdated and fail to handle constantly evolving tweets. In an effort to mitigate this issue, we address the problem of vaccination opinion mining by learning disentangled stance and aspect vectors of tweets in order to cluster tweets along the aspect axis.

Our proposed model, called Disentangled Opinion Clustering (DOC), is shown in Figure 1. It is trained in two steps. In unsupervised learning (Figure 1(a)), a tweet is fed into an autoencoder with DeBERTa as both the encoder and the decoder to learn the latent sentence vector z. Here, each tweet is mapped to two embeddings, the context embedding u, which encodes the stance label information and the aspect embedding u_a which captures the aspect information. Under unsupervised learning, these two embeddings are not distinguished. Together with the hidden representation of the input text, H, they are mapped to the latent sentence vector z by cross-attention. As the autoencoder can be trained on a large-scale unannotated tweets relating to vaccination, it is expected that z would capture the vaccine-related topics.

Then in the second step of supervised learning (Figure 1(b)), the DeBERTa-based autoencoder is fine-tuned to learn the latent stance vector z_s and the latent aspect vector z_a using the tweet-level annotated stance label and aspect text span (or the argumentative pattern ‘vaccine safety’ in Figure 1(b)) as the inductive bias. Here, the latent stance vector z_s is derived from u_s. It is expected that z_s can be used to predict the stance label. On the other hand, the latent aspect vector z_a is derived from u_a only and it can be used to generate the SBERT-encoded aspect text span. Both z_s and z_a, together with the hidden representation of the input text H, are used to reconstruct the original text through the DeBERTa decoder. The training instances are organized in pairs since we use the idea of swapped autoencoder (shown in Figure 1(c)) to swap the aspect embedding of one tweet with that of another if both discuss the same aspect. The resulting latent vector can still be used to reconstruct the original tweet. In what follows, we describe the two steps, unsupervised and supervised learning, in detail.

Unsupervised Learning of Sentence Representation Due to the versatility of PLMs, sentence representations are usually derived directly from contextualised representations generated by the PLMs. However, as has been previously discussed in Montero et al. (2021), sentence representations derived in this way cannot guarantee reliable reconstruction of the input text. Partly inspired by the use of autoencoder for sentence representation learning as in (Montero et al., 2021), we adopt the autoencoder...
The Pfizer vaccine is safe

\[ z_A \]  DeBERTa Encoder

Cross Attention

DeBERTa Decoder

\[ 2 \]

Unsupervised Learning

The Pfizer [mask] is safe

\[ z_A \]  DeBERTa Encoder

Cross Attention

DeBERTa Decoder

\[ 2 \]

Supervised Learning

\[ y_A \]  Softmax

\[ A \]  MLP

\[ a \]

Swapped AutoEncoder

The Pfizer [mask] is safe

\[ z_B \]  DeBERTa Encoder

Cross Attention

DeBERTa Decoder

\[ 1 \]

\[ b \]

\[ H \]

architecture to initially guide the sentence representation learning by fine-tuning it on vaccination tweets. Rather than RoBERTa (Liu et al., 2019), we adopt DeBERTa, a variant of BERT in which each word is represented using two vectors encoding its content and position. The attention weight of a word pair is computed as a sum of four attention scores calculated from different directions based on their content/position vectors, i.e., content-to-content, content-to-position, position-to-content, and position-to-position. Instead of representing each word by a content vector and a position vector, we modify DeBERTa by representing an input sentence using two vectors, a context embedding \( u_s \) encoding its stance label information and an aspect embedding \( u_a \) encoding its aspect information. We will discuss later in this section how to perform disentangled representation learning with \( u_s \) and \( u_a \). During the unsupervised learning stage, we do not distinguish between \( u_s \) and \( u_a \) and simply use \( u = [u_s, u_a] \) to denote them.

More specifically, we train the autoencoder on an unannotated Twitter corpus with the masked token prediction as the training objective. The encoder applies the multi-head attention to clamp the hidden representations of the top layer of the pre-trained transformer. If we use \( H \) to denote the hidden representations, the multi-head attention can be expressed as:

\[
\text{head}_i = \text{softmax} \left( \frac{uW_Q (HW_K)^T}{\sqrt{d_H}} \right) HW_V, \quad (1)
\]

\[
z = [\text{head}_1, \text{head}_2, \ldots, \text{head}_d] W_O, \quad (2)
\]

where \( H \in \mathbb{R}^{n \times d_H} \), \( W_Q \in \mathbb{R}^{2d_H \times d_K} \), \( W_K \in \mathbb{R}^{d_H \times d_K} \), \( W_V \in \mathbb{R}^{d_H \times d_V} \), \( \text{head}_i \in \mathbb{R}^{d_V} \) and \( W_O \in \mathbb{R}^{hd_V \times d_a} \). \( u \in \mathbb{R}^{2d_H} \) is generated from a fully-connected layer over the hidden vectors. The bottleneck representation \( z \) is supposed to encode the semantics of the whole sentence.

The transformer decoder comprises \( n \) layers of cross-attention such that the output of the previous layer is processed by a gating mechanism (Hochreiter and Schmidhuber, 1997). The recurrence is repeated \( n \) times to reconstruct the input, where \( n \) denotes the token length of the input text.

Injecting Inductive Biases by Disentangled Attention Recent work on disentanglement learning suggested unsupervised disentanglement is impossible without inductive bias (Locatello et al., 2020b). In the datasets used in our experiments, there are a small number of labelled tweets. We can use the tweet-level stance labels and the annotated aspect text spans as inductive bias. Here, the disentangled attention of DeBERTa is utilized to mingle different factors. Assuming each sentence is mapped to two vectors, the context vector \( u_s \) encoding its stance label information and the aspect vector \( u_a \) encoding its aspect information, we can then map \( u_s \) to a latent stance vector \( z_s \).
which can be used to predict the stance label, and map \( u_a \) to a latent aspect vector \( z_a \) which can be used to reconstruct the aspect text span. We use the cross-attention between \( u_s \) and \( u_a \) to reconstruct the original input sentence.

**Stance Classification** Let \( h_{CLS} \) denote the hidden representation of the \([CLS]\) token, the stance bias is injected by classification over the stance categories:

\[
z_a = \text{softmax} \left( \frac{u_s W_q a (h_{CLS} W_k a)^\top}{\sqrt{d_H}} \right) h_{CLS} W_v, \quad (3)
\]

\[
y_a = \text{softmax}(z_a W), \quad L_s = -\hat{y}_a^{(y)} \log \hat{y}_a^{(y)}. \quad (4)
\]

Essentially, we use \( u_s \) as query and \( h_{CLS} \) as key and value to derive \( z_s \), which is subsequently fed to a softmax layer to predict a stance label \( \hat{y}_a \). The objective function is a cross-entropy loss between the true and the predicted labels.

**Aspect Text Span Reconstruction** We assume \( u_s \) encoding the sentence-level aspect information and use self-attention to derive the latent aspect representation \( z_a \). To reconstruct the aspect text span from \( z_a \), we use the embedding generated by SBERT (Reimers and Gurevych, 2019) as the targeted aspect span, since SBERT has been empirically shown achieving the state-of-the-art on Semantic Textual Similarity tasks. Those clustering-friendly representations, if they encode the argumentative patterns or aspect spans alone, are strong inductive biases in the axis of aspects.

Specifically, the sentence embedding of the aspect expression is generated by a Gaussian MLP decoder (Kingma and Welling, 2014):

\[
z_a = \text{softmax} \left( \frac{u_s W_q a (u_a W_k a)^\top}{\sqrt{d_H}} \right) u_a W_v, \quad (5)
\]

\[
L_a = -\log N(y_a; \text{MLP}_p(z_a), \text{MLP}_o(z_a) I), \quad (6)
\]

where \( x_a \) denotes the aspect text span in the original input sentence, \( y_a \) is the ground-truth aspect text span embedding produced by \( y_a = \text{SBERT}(x_a) \), whose value is used for computing the Gaussian negative log-likelihood loss\(^2\).

**Input Text Reconstruction** To reconstruct the original input text, we need to make use of both the latent stance vector \( z_s \) and the latent aspect vector \( z_a \). Here we use the cross-attention of these two vectors to derive the content vector \( z_c \).

\[
Q_c = u W_q c, \quad K_c = H W_k c, \quad V_c = H W_v c,
\]

\[
Q_s = u_s W_q s, \quad K_s = u_s W_k s, \quad Q_a = u_a W_q a, \quad K_a = u_a W_k a.
\]

\[
a_j = Q_c K_j^\top + Q_c K_s^\top + Q_a K_c^\top + Q_c K_a^\top + K_c^\top + K_s^\top + K_a^\top
\]

\[
\text{head}_i = \text{softmax} \left( \frac{a}{\sqrt{d_H}} \right) H W_v, \quad z_c = [\text{head}_1, \text{head}_2, \ldots, \text{head}_n] W_O, \quad (7)
\]

where \( u = [u_s, u_a] \), \( a_j \) is the \( j \)-th element of \( a \), and \( K_j^\top \) represents the \( j \)-th row of \( K_c \). The resulting \( z_c \) is the content representation for reconstructing the original sentence.

**Disentanglement of Aspect and Stance** Although the inductive biases, i.e., the tweet-level stance label and the annotated aspect span, are used to learn the latent stance vectors \( z_s \) and the aspect vectors \( z_a \), there could still be possible dependencies between the two latent variables. To further the disentanglement, we propose to swap the learned aspect embeddings of two tweets discussing the same aspect in Siamese networks. We draw inspiration from the Swapping Autoencoder (Park et al., 2020) where a constituent vector of a Generative Adversarial Network (GAN) is swapped with that produced by another image. The original swapping autoencoder was designed for image editing and required a patch discriminator with texture cropping to the corresponding disentangled factors with the desired properties. In our scenario, such alignment is instead induced by tweets discussing the same aspect.

We create pairs of tweets by permutations within the same aspect group \((x^A, x^B)\) \(A,B \in G_k, A \neq B\). Here, by abuse of notation, we use \( k \) to denote the \( k \)-th aspect group, \( G_k \). The groups are identified by tweets with the same aspect label, regardless of their stances. We sketch the structure of pair-wised training in Figure 1(c). The tweets are organized in pairs and a bottleneck representation is obtained for each tweet:

\[
z^A = \text{enc}(x^A), \quad z^B = \text{enc}(x^B). \quad (8)
\]

We would like \( z^A \) to disentangle into latent factors, i.e., the variation in a factor of \( z^A \) is associated with a change in \( x^A \) (Locatello et al., 2020a). Unlike the majority of works (Zhang et al., 2021d) that directly splits \( z^A \) in the latent space, we assume that the entangled vector is decomposed by a causal network. We train a vector \( u = [u_s, u_a] \) to trigger the activation of the networks (i.e., the self-attentions in Eq. 3-Eq. 7). The outputs of the networks are

---

independent components that encode the desiderata. If \( z_a \) and \( z_u \) are parameterized independent components triggered by \( u_u \) and \( u_s \), respectively, the substitution of \( u_u^B \) with \( u_u^A \) can be regarded as soft exchanges between \( z_u^A \) and \( z_u^B \).

We thus substitute \( u_u^B \) with \( u_u^A \) to cause changes in \( z_u^B \). This substitution will also be reflected by changes in \( z_u^A \). In practice, we train on all permutations with the same aspect group, regardless of the stance. The reconstruction loss for each latent factor (i.e., stance and aspect) is calculated once to cause changes in \( z_u^A \).

Formally, the swapping autoencoder presented in Figure 1(c) can be expressed as

\[
Q_u^B = u_u^B W_{q_u,s}, \quad K_u^B = u_u^B W_{k_u,s}, \\
Q_u^A = u_u^A W_{q_u,a}, \quad K_u^A = u_u^A W_{k_u,a}, \\
a_j = Q_u^B a_j^T + Q_u^A a_j^T + K_u^B a_j^T + Q_u^B a_j^T + K_u^A a_j^T + Q_u^A a_j^T, \\
head_i = \text{softmax} \left( \frac{a_j}{\sqrt{d_H}} \right) HW_{u,c}, \\
z_u^B = [\text{head}_1, \text{head}_2, \ldots, \text{head}_n] W_O, \\
z_u^B = \text{softmax} \left( \frac{u_u^B W_{q_u,a}(K_u^A)^T}{\sqrt{d_H}} \right) V_{CLS}, \\
z_u^B = \text{softmax} \left( \frac{Q_u^A (K_u^A)^T}{\sqrt{d_H}} \right) u_u^A W_{v,a},
\]

where \( z_u^B \) is input to the decoder for the reconstruction of \( z_u^B \). Note that the above equations are specially used in the swapping autoencoder for the computation of \( z_u^B \). If there is no substitution in the latent space, the above equations will not be calculated. Given \( L_u^B = \text{dec}(z_u^B) \), the final objective function is written as

\[
L = L_u^A + \lambda_a L_u^A + \lambda_d L_u^A + \lambda_B L_u^B,
\]

where \( \lambda_a \), \( \lambda_B \) are hyper-parameters controlling the importance of each desirable property. In our experiments, we choose \( \lambda_a = \lambda_B = 1 \) and \( \lambda_B = 0.5 \).

### 4 Experiments

**Datasets** We conduct our experimental evaluation on two publicly available Twitter datasets about the Covid-19 vaccination: the Covid-Moral-Foundation (CMF) (Pacheco et al., 2022) and the Vaccination Attitude Detection (VAD) corpus (Zhu et al., 2022). CMF is a tweet dataset focused on the Covid-19 vaccine debates, where each tweet is assigned an argumentative pattern. VAD consists of 8 aspect categories further refined by vaccine bands. Similar to the argumentative pattern in the CMF dataset, each tweet is characterised by a text span indicating its aspect. The dataset statistics are reported in Table 1, with examples shown in A.1. The train/test split follows 4 : 1. For the unsupervised pre-training of sentence bottleneck representations, we combine the unlabelled Covid-19 datasets from both CMF\(^3\) and VAD\(^4\) repositories. The final dataset consists of 4.37 million tweets.

**Baselines** We employ 5 baseline approaches: SBERT\(^5\), AutoBot\(^6\), DS-Clustering, VADet, and SCCL\(^7\), of which SBERT and AutoBot are out-of-the-box sentence embedding generators. VADet is specialised to learn disentangled representations. However, it is noteworthy that even though it employs DEC (Xie et al., 2016), the resulting representations are unsuitable for distance-based clustering. SCCL performs joint representation learning and document clustering. DS-Clustering is a pipeline approach that predicts a text span and employs SBERT to generate an aspect embedding. For clustering-friendly representation learning methods, we examine their performance using \( k \)-means and \( k \)-medoids (Leonard and Peter, 1990), and

### Table 1: Dataset statistics of CMF and VAD. We list the number of pro-vaccine, anti-vaccine and neutral tweets in each group.

<table>
<thead>
<tr>
<th>Aspect Group</th>
<th>Pro-Vax</th>
<th>Anti-Vax</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMF</td>
<td>400</td>
<td>84</td>
<td>36</td>
</tr>
<tr>
<td>Personal Experience</td>
<td>381</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Vaccines Save Lives</td>
<td>12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(Adverse) Side Effects</td>
<td>179</td>
<td>256</td>
<td>63</td>
</tr>
<tr>
<td>Immunity Level</td>
<td>433</td>
<td>113</td>
<td>52</td>
</tr>
<tr>
<td>Economic Effects</td>
<td>23</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Personal Freedom</td>
<td>5</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Moral Attitudes</td>
<td>5</td>
<td>43</td>
<td>2</td>
</tr>
</tbody>
</table>

\(^3\)https://gitlab.com/mlpacheco/covid-moral-foundations
\(^4\)https://github.com/somethingx1202/VADet
\(^5\)https://github.com/UKPLab/sentence-transformers
\(^6\)https://github.com/ivanmontero/autobot
\(^7\)https://github.com/amazon-research/sccl
the Agglomerative Hierarchical Clustering (AHC). The comparison involves three tasks: tweet clustering based on aspect categories (intra- and cross-datasets), and tweet-level stance classification. For stance classification, we employ RoBERTa and DeBERTa, and use their averaged embeddings for clustering.

**Evaluation Metrics** First, we use Clustering Accuracy (CA) and Normalized Mutual Information (NMI) to evaluate the quality of clusters in line with (Shaham et al., 2018; Tao et al., 2021). NMI is defined as $\text{NMI} = \frac{(2 \times \text{I}(y; \hat{y}))}{(\text{H}(y) + \text{H}(\hat{y}))}$, where $\text{I}(y; \hat{y})$ denotes the mutual information between the ground-truth labels and the predicted labels, $\text{H}(\cdot)$ denotes their entropy. Then we employ BERTScore (Zhang et al., 2020) to evaluate the performance of models in clustering in the absence of ground-truth cluster labels. BERTScore is a successor of Cosine Similarity (John et al., 2019) that measures the sentence distance by calculating the cross distance between their corresponding word embeddings. We follow Bilal et al. (2021) to compute the averaged BERTScore as

$$\text{AvgBS} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\binom{|G_k|}{2}} \sum_{i<j \in G_k} \text{BS}(\text{tweet}_i, \text{tweet}_j),$$

where $|G_k|$ is the size of the $k$-th group or cluster. We report the average performance for all the models. As a quantitative evaluation metric for disentanglement, we use the Mean Correlation Coefficient (MCC). We refer the readers to A.3 for qualitative results.

**Clustering-Friendly Representation** We first show the advantages of disentangled representations in clustering. With the representations obtained from SBERT and AutoBot, we employ $k$-means to perform clustering. Since the similarity between sentences in SBERT is measured by cosine similarity which is less favorable for $k$-means algorithm, we also use $k$-medoids to ensure a fair comparison. The other baseline approaches are run with their default settings. We assign the aspect labels to the predicted clusters with the optimal permutation such that the permutation of $\{1, \ldots, K\}$ yields the highest accuracy score, where $K$ denotes the total number of clusters. For the CMF dataset, we set $K = 7$, and on VAD $K = 8$.

Table 2 lists the performance of baseline methods on all the tasks and datasets. We see consistent improvements across all the evaluation metrics using our proposed DOC. When compared with end-to-end methods (i.e., VADet and SCCL) whose intermediate representations cannot be used to calculate a distance, the disparity depends on DOC’s clustering approaches employed. On CMF, VADet outperforms SCCL. But DOC gives superior performance overall regardless of the clustering approaches used, showing the flexibility of the DOC representations. In comparisons against representation learning methods, DOC takes the lead as long as it is attached with competent clustering algorithms. This shows the benefit of clustering with disentangled representations since the clustering algorithm will no longer obfuscate the stance polarities and the aspect categories. DOC achieves higher scores on the VAD dataset compared to CMF, with more prominent improvement over the baselines, which may be credited to the increased size of the dataset. When DOC is evaluated with different clustering algorithms, $k$-medoids excels on CMF, while AHC outperforms the others on VAD, showing that cosine similarity is more appropriate for distance calculation since the $k$-means algorithm relies on Euclidean distance.

**Cross-Dataset Evaluation** In this context, the most interesting property of clustering-friendly representations is their ability to perform clustering in novel datasets whose categories are unknown in advance. To assess this, we use the models trained on CMF to perform clustering on VAD, and repeat the process vice versa. We specify the number of
Table 3: Cross-dataset evaluation results. Each representation learning model is listed with the most performant clustering method.

<table>
<thead>
<tr>
<th>Models</th>
<th>VAD → CMF</th>
<th>CMF → VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CA</td>
<td>NMI</td>
</tr>
<tr>
<td>SBERT-AHC</td>
<td>51.6</td>
<td>49.8</td>
</tr>
<tr>
<td>AutoBot-(k)-medoids</td>
<td>53.1</td>
<td>50.6</td>
</tr>
<tr>
<td>DS-C-(k)-medoids</td>
<td>54.1</td>
<td>51.2</td>
</tr>
<tr>
<td>VADet</td>
<td>53.5</td>
<td>50.1</td>
</tr>
<tr>
<td>SCCL</td>
<td>48.6</td>
<td>47.0</td>
</tr>
<tr>
<td>DOC-(k)-medoids</td>
<td>55.3</td>
<td>51.9</td>
</tr>
<tr>
<td>DOC-AHC</td>
<td>53.5</td>
<td>50.4</td>
</tr>
</tbody>
</table>

Table 4: Stance classification results.

<table>
<thead>
<tr>
<th>Models</th>
<th>CMF</th>
<th>VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Micro F1</td>
<td>Macro F1</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>72.3±0.5</td>
<td>71.2±0.4</td>
</tr>
<tr>
<td>DeBERTa</td>
<td>74.0±0.6</td>
<td>73.5±0.6</td>
</tr>
<tr>
<td>DOC-AHC</td>
<td>73.5±0.6</td>
<td>72.7±0.6</td>
</tr>
</tbody>
</table>

Table 5: Ablation study on removal of components and choices of context vectors.

<table>
<thead>
<tr>
<th>Model</th>
<th>CMF</th>
<th>VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOC-(k)-means</td>
<td>51.7</td>
<td>18.5</td>
</tr>
<tr>
<td>w/o pre-trained LM</td>
<td>43.3</td>
<td>16.2</td>
</tr>
<tr>
<td>w/o inductive bias</td>
<td>50.0</td>
<td>18.0</td>
</tr>
<tr>
<td>w/o swapped codes</td>
<td>50.8</td>
<td>17.8</td>
</tr>
</tbody>
</table>

Figure 2: Boxplots of MCC for all representation learning models, over the 5 runs. The representations are used for \(k\)-means clustering in the Euclidean space. A high MCC score indicates the strong correlation between \(\text{dist}(z_a, z_k^a)\) and \(z_a \in G_k\).

**Stance Classification** We report in Table 4 the results of DOC, RoBERTa and DeBERTa. For DOC, we only report DOC-AHC since stance labels are by-products of clustering-friendly representations. We see the DOC performance on CMF close to that of DeBERTa, and that the improvement on VAD is marginal. This may be attributed to the absence of the swapping operation on \(z_a\), and therefore the stance latent vector may contain other semantics or noise. Nevertheless, DOC is still preferred over DeBERTa considering its significant performance gain over DeBERTa on aspect clustering.

**Ablations Study** We study the effects by taking away components of different functionality in disentanglement, and experiment with different choices of context vectors, i.e., \(u_s\) and \(u_a\). The results are shown in Table 5. We see a significant performance drop without loading the pre-trained weights for the language model. The removal of inductive biases and the swapped autoencoder both hamper the clustering of the model across the metrics. The performance gap is more obvious without the inductive bias, which we attribute to the weaker supervision induced by swapping the latent codes. Ablating choices of context vectors show the superiority of the MLP strategy. In contrast, the performance of the context vector generated by mean pooling is rather poor. It shows that the context vector produced by mean-pooling can hardly trigger the disentanglement of the hidden semantics.
tween the aspect vector and the centroid of cluster \( k \) and \( Y \) (i.e., the dichotomous variable indicating whether it belongs to or not belongs to group \( k \) in groundtruth) is chosen to measure the isometry between \( z_a \) and \( k \). Notice that we specify \( dist \) as Euclidean Distance here. However, isometry does not hinge on the Euclidean Distance, and it could be easily substituted with Cosine Similarity, in which case the mean is no longer the best estimation for the cluster center and would be replaced by the medoid of cluster \( k \). The clustering method would be \( k \)-medoids accordingly.

For each cluster \( k \in \{1, 2, \ldots, K\} \), we calculate the correlation coefficient between \( dist(z_a, \bar{z}_a^k) \) and \( Y \). We then obtain MCC by averaging the correlation coefficients. A high MCC indicates that the group identity of a data point is closely associated with the geometric position of its \( z_a \) in the latent space, which means that \( z_a \) captures the group information. The results are shown in Figure 2. We observe consistent improvement over the sentence representation models. DS-Clustering is able to encode tweets into aspect embeddings. Nevertheless, its distance between aspect latent vectors is a weaker indicator for group partition compared with that of DOC, suggesting that \( z_a \) discovered by DOC better captures the difference between aspects.

5 Conclusion

In this work, we introduced DOC, a **Disentangled Opinion Clustering** model for vaccination opinion mining from social media. DOC is able to disentangle users’ stances from opinions via a disentangling attention mechanism and a swap-autoencoder. It was designed to process unseen aspect categories thanks to the clustering approach, leveraging clustering-friendly representations induced by out-of-the-box Sentence-BERT encodings and the disentangling mechanisms. A thorough experimental assessment demonstrated the benefit of the disentangling mechanism on the quality of aspect-based clusters and the generalization capability across datasets with different aspect categories outperforming existing approaches in terms of generalisation and coherence of the generated clusters.

6 Limitations

There are a few limitations we would like to address. First of all, the number of clusters needs manual configuration. This is a limitation of the clustering algorithms (Xie et al., 2016) since we need to set a threshold for convergence, which consequentially pinpoints \( k \). An expedient alternative is to analyse the dataset for the realistic settings or probe into \( k \) for the optimal setup, which is, however, beyond the scope of this paper. Another limitation is the pre-requisite for millions of unannotated data. The autoencoder needs enormous data to learn bottleneck representations. Its performance would be hindered without access to abundant corpora. Lastly, the performance of the acquired clustering-friendly representations depends on the similarity metric chosen. Efforts need to be made to find the best option, whether it is Euclidean distance or cosine similarity etc.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council (grant no. EP/T017112/1, EP/V048597/1), YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (EP/V020579/2). This work was conducted on the UKRI/EPSRC HPC platform, Avon, hosted in the University of Warwick’s Scientific Computing Group.

References


Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. 2016. Disentangling factors of variation in deep representation using adversarial training. In *Advances in
Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01):6826–6833.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022, WWW ’22*, page 3143–3152, New York, NY, USA. Association for Computing Machinery.


Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed Elsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. 2021. COVID-19 literature knowledge graph construction and drug repurposing report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77, Online. Association for Computational Linguistics.


Wei Zhang, Chao Dong, Jianhua Yin, and Jianyong Wang. 2021c. Attentive representation learning with adversarial training for short text clustering. IEEE Transactions on Knowledge and Data Engineering.


A Appendix

A.1 Dataset Details

In this section, we provide a detailed analysis of the dataset instances.

In the Covid-Moral-Foundation (CMF) dataset, each tweet is associated with a pre-defined and manually annotated argumentative pattern. The annotated tweets are categorized by moral foundations that can be regarded as coarse aspects distilled from argumentative patterns. Each moral foundation is associated with two polarities (e.g., care/harm), and is treated as the group label of a cluster of tweets. The polarity is given by the vaccination stance label. In the example in Table A1, ‘The vaccine is safe’ is the argumentative pattern, while ‘Care/Harm’ is the categorical label denoting the aspect group. An exhaustive list to the argumentative patterns can be found in the original paper of Pacheco et al. (2022).

In the Vaccination Attitude Detection (VAD), a training instance comprises a stance label, a categorical aspect label and an aspect text span. For example, Table A1 shows the tweet ‘Study reports Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19,’ is annotated with the text span ‘Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19’, and its aspect belongs to the aspect category ‘Immunity Level’.

A.2 Training Details

We experiment with a pre-trained DeBERTa\(^8\) base model. The hidden size is \(d_H = 768\). We set both \(d_V\) and \(d_K = 768\), and \(d_a = 1024\). The learning rate is initialised with \(\eta = 3e − 5\) and the number of epochs is 10. We use Linear Warmup to enforce the triangular learning rate.

We train the model with two Titan RTX graphics cards on a station of an Intel(R) Xeon(R) W-2245 CPU. The training process takes less than 9 hours, with the inference time under 30 minutes.

A.3 Additional Results

Clustering with Different Latent Vectors

We experiment clustering using the disentangled aspect vectors \(z_a\) or the content vectors \(z\) (i.e., without the disentanglement of aspects and stances) on both CMF and VAD datasets, and have the detailed results reported in Table A2. It can be observed that using the disentangled aspect vectors for clustering gives better results compared to using the content vectors, regardless of the clustering approaches used. On CMF, the best results are obtained using \(k\)-medoids, while on VAD, similar results are obtained using either \(k\)-medoids or AHC.

\(^8\)https://huggingface.co/docs/transformers/model_doc/deberta-v2
Vaccine decreases your chances of getting severe life-threat.
The vaccine is safe
Care/Harm

There is no way someone can tell me that the COVID vaccine does not cause harm to pregnant women.
The covid vaccine is harmful for pregnant women and kids
Care/Harm

The tyranny is not locking down and not using the vaccine to appease the crazies who think it’s oppression.
The vaccine mandate is not oppression because it will help to end this pandemic
Liberty/Oppression

Table A1: Training examples of CMF and VAD. In CMF, Argumentative Patterns are pre-defined phrases indicating an aspect. In VAD, aspect spans are text sub-sequence of the annotated tweets.

<table>
<thead>
<tr>
<th>Latent Vector</th>
<th>CMF CA</th>
<th>AvgBS</th>
<th>VAD CA</th>
<th>AvgBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOC-k-means-za</td>
<td>51.7</td>
<td>18.5</td>
<td>64.2</td>
<td>20.3</td>
</tr>
<tr>
<td>DOC-k-means-z</td>
<td>48.3</td>
<td>17.5</td>
<td>60.7</td>
<td>18.7</td>
</tr>
<tr>
<td>DOC-k-medoids-za</td>
<td>54.2</td>
<td>20.7</td>
<td>66.7</td>
<td>21.4</td>
</tr>
<tr>
<td>DOC-k-medoids-z</td>
<td>50.8</td>
<td>18.0</td>
<td>61.4</td>
<td>18.9</td>
</tr>
<tr>
<td>DOC-AHC-za</td>
<td>52.5</td>
<td>19.1</td>
<td>66.7</td>
<td>22.8</td>
</tr>
<tr>
<td>DOC-AHC-z</td>
<td>49.2</td>
<td>17.8</td>
<td>61.9</td>
<td>19.0</td>
</tr>
</tbody>
</table>

Table A2: Clustering accuracy and average BERTScore with different latent vectors.

Color Mappings in Visualisation
We illustrate in Figure A2 the color mapping from t-SNE plots to the true aspect category labels. It is shown that the vectors are more separated and their grouping aligns closer to the ground-truth labels when they are clustered on the space of $z_a$, indicating that such latent vectors provide strong distance-based discrimination among groups in the Euclidean space, as has been used as a distance metric in the t-SNE algorithm. We also experiment with cosine-similarity metric for $k$-medoids and the results have been reported in the Experiments section.
Figure A2: t-SNE plots on CMF and VAD. Each dot is a tweet encoded using either the disentangled aspect vector $z_a$ (left subfigure) or the latent content vector $z$ (right subfigure). Different colors indicate the true aspect category labels.
ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
  Section 7

- A2. Did you discuss any potential risks of your work?
  Our work does not introduce a novel dataset.

- A3. Do the abstract and introduction summarize the paper's main claims?
  Section 1

- A4. Have you used AI writing assistants when working on this paper?
  Left blank.

B ✓ Did you use or create scientific artifacts?

- B1. Did you cite the creators of artifacts you used?
  Section 3

- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
  Section 4

- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
  Our work does not create new datasets. Our model is not designed for specific purposes.

- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
  The data are anonymous, as stated in their publications of origin. We double-checked the datasets and can confirm that they are anonymous.

- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
  Section 4

- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
  Section 4

C ✓ Did you run computational experiments?

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
  Appendix A2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, Section 7, Appendix A2

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?
No response.

D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.