Disentangling Aspect and Stance via a Siamese Autoencoder for Aspect Clustering of Vaccination Opinions

Lixing Zhu[†], **Runcong Zhao**[†], **Gabriele Pergola**[‡], **Yulan He**[†,‡,§]
[†]Department of Computer Science, University of Warwick, UK
[‡]Department of Informatics, King's College London, UK
[§]The Alan Turing Institute, UK
{lixing.zhu,runcong.zhao,yulan.he}@kcl.ac.uk
gabriele.pergola.1@warwick.ac.uk

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 (Hussain 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-

Findings of the Association for Computational Linguistics: ACL 2023, pages 1827–1842
July 9-14, 2023 ©2023 Association for Computational Linguistics

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 perturbated 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).

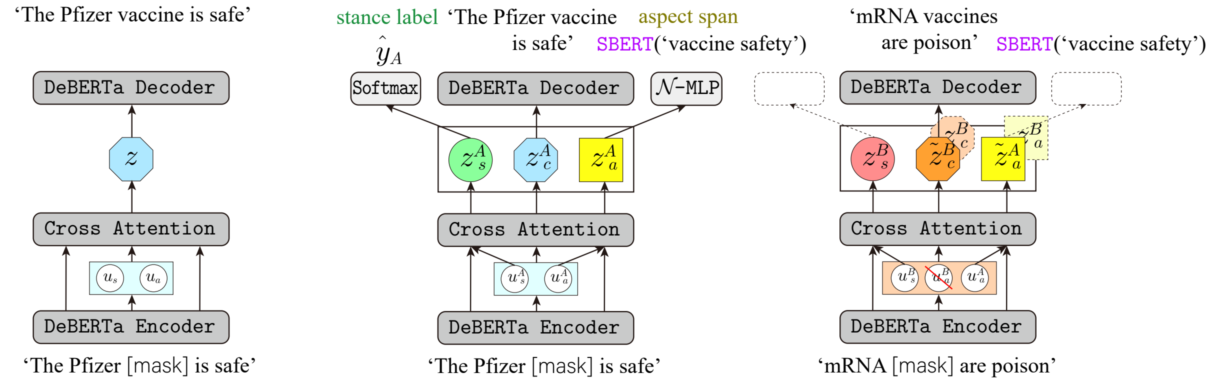
3 Methodology

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 $\boldsymbol{z}$. Here, each tweet is mapped to two embeddings, the context embedding $\boldsymbol{u}_s$ which encodes the stance label information and the aspect embedding $\boldsymbol{u}_a$ 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 $\boldsymbol{z}$ by cross-attention. As the autoencoder can be trained on a large-scale unannotated tweets relating to vaccination, it is expected that $\boldsymbol{z}$ 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 $\boldsymbol{z}_s$ and the latent aspect vector $\boldsymbol{z}_a$ 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 $\boldsymbol{z}_s$ is derived from $\boldsymbol{u}_s$. It is expected that $\boldsymbol{z}_s$ can be used to predict the stance label. On the other hand, the latent aspect vector $\boldsymbol{z}_a$ is derived from $\boldsymbol{u}_a$ only and it can be used to generate the SBERT-encoded aspect text span. Both $\boldsymbol{z}_s$ and $\boldsymbol{z}_a$, 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



(a) Unsupervised Learning (b) Supervised Learning (c) Swapped AutoEncoder

Figure 1: Disentangled Opinion Clustering (DOC) Model. **(a) Unsupervised learning.** A tweet is fed into an autoencoder with DeBERTa as both the encoder and decoder to learn the latent sentence vector $\boldsymbol{z}$; **(b) Supervised learning.** The DeBERTa-based autoencoder is fine-tuned to learn the latent stance vector $\boldsymbol{z}_s$ and the latent aspect vector $\boldsymbol{z}_a$ using the tweet-level annotated stance label and aspect text span (or the argumentative pattern ‘*vaccine safety*’ for the input tweet) as the inductive bias; **(c) Swapping autoencoder.** To enable a better disentanglement of $\boldsymbol{z}_s$ and $\boldsymbol{z}_a$, for the two tweets discussing the same aspect but with different stance labels, tweet B’s aspect embedding $\boldsymbol{u}_a^B$ is replaced by the tweet A’s aspect embedding $\boldsymbol{u}_a^A$. As the two tweets discuss the same aspect, their aspect embeddings are expected to be similar. As such, we can still reconstruct tweet B using the latent content vector $\boldsymbol{z}_c^B$ derived from the swapped aspect embedding. Note that (b) and (c) are learned simultaneously.

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 $\boldsymbol{u}_s$ encoding its stance label information and an aspect embedding $\boldsymbol{u}_a$ encoding its aspect information. We will discuss later in this section how to perform disentangled representation learning with $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$. During the unsupervised learning stage, we do not distinguish between $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$ and simply use $\boldsymbol{u} = [\boldsymbol{u}_s, \boldsymbol{u}_a]$ 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:

$$\text{head}_i = \text{softmax}\left(\frac{\mathbf{u}W_Q(HW_K)^\top}{\sqrt{d_H}}\right)HW_V, \quad (1)$$

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

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 $W_O \in \mathbb{R}^{hd_V \times d_\mathbf{z}}$. $\mathbf{u} \in \mathbb{R}^{2d_H}$ is generated from a fully-connected layer over the hidden vectors. The bottleneck representation $\mathbf{z}$ 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 $\boldsymbol{u}_s$ encoding its stance label information and the aspect vector $\boldsymbol{u}_a$ encoding its aspect information, we can then map $\boldsymbol{u}_s$ to a latent stance vector $\boldsymbol{z}_s$

which can be used to predict the stance label, and map $\boldsymbol{u}_a$ to a latent aspect vector $\boldsymbol{z}_a$ which can be used to reconstruct the aspect text span. We use the cross-attention between $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$ to reconstruct the original input sentence.

Stance Classification Let $\mathbf{h}_{\mathsf{CLS}}$ denote the hidden representation of the `[CLS]` token, the stance bias is injected by classification over the stance categories:

$$\mathbf{z}_{\boldsymbol{s}} = \mathrm{softmax}\left(\frac{\mathbf{u}_{\boldsymbol{s}} W_{\boldsymbol{q},\boldsymbol{s}}(\mathbf{h}_{\mathsf{CLS}} W_{\boldsymbol{k},\mathsf{CLS}})^{\top}}{\sqrt{d_H}}\right)\mathbf{h}_{\mathsf{CLS}} W_{\boldsymbol{v},\mathsf{CLS}}, \quad (3)$$

$$\hat{y}_s = \mathrm{softmax}(\mathbf{z}_{\boldsymbol{s}} W), \quad \mathcal{L}_s = -y_s^{(i)} \log \hat{y}_s^{(i)}. \quad (4)$$

Essentially, we use $\boldsymbol{u}_s$ as query and $\mathbf{h}_{\mathsf{CLS}}$ as key and value to derive $\boldsymbol{z}_s$, which is subsequently fed to a softmax layer to predict a stance label $\hat{y}_s$. The objective function is a cross-entropy loss between the true and the predicted labels.

Aspect Text Span Reconstruction We assume $\boldsymbol{u}_a$ encoding the sentence-level aspect information and use self-attention to derive the latent aspect representation $\boldsymbol{z}_a$. To reconstruct the aspect text span from $\boldsymbol{z}_a$, 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):

$$\mathbf{z}_a = \mathrm{softmax}\left(\frac{\mathbf{u}_a W_{\boldsymbol{q},\boldsymbol{a}}(\mathbf{u}_a W_{\boldsymbol{k},\boldsymbol{a}})^{\top}}{\sqrt{d_H}}\right)\mathbf{u}_a W_{\boldsymbol{v},\boldsymbol{a}}, \quad (5)$$

$$\mathcal{L}_a = -\log \mathcal{N}(\mathbf{y}_a; \mathrm{MLP}_{\mu}(\mathbf{z}_{\boldsymbol{a}}), \mathrm{MLP}_{\sigma}(\mathbf{z}_{\boldsymbol{a}})\mathbf{I}), \quad (6)$$

where $\boldsymbol{x}_a$ denotes the aspect text span in the original input sentence, $\mathbf{y}_a$ is the ground-truth aspect text span embedding produced by $\mathbf{y}_a = \mathsf{SBERT}(\boldsymbol{x}_a)$, 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 $\boldsymbol{z}_s$ and the latent aspect vector $\boldsymbol{z}_a$. Here we use the cross attention of these two vectors to derive the content vector $\boldsymbol{z}_c$.

$$\begin{gathered} Q_{\boldsymbol{c}} = \mathbf{u} W_{\boldsymbol{q},\boldsymbol{c}}, \quad K_{\boldsymbol{c}} = H W_{\boldsymbol{k},\boldsymbol{c}}, \quad V_{\boldsymbol{c}} = H W_{\boldsymbol{v},\boldsymbol{c}}, \\ Q_{\boldsymbol{s}} = \mathbf{u}_{\boldsymbol{s}} W_{\boldsymbol{q},\boldsymbol{s}}, \quad K_{\boldsymbol{s}} = \mathbf{u}_{\boldsymbol{s}} W_{\boldsymbol{k},\boldsymbol{s}}, \\ Q_{\boldsymbol{a}} = \mathbf{u}_{\boldsymbol{a}} W_{\boldsymbol{q},\boldsymbol{a}}, \quad K_{\boldsymbol{a}} = \mathbf{u}_{\boldsymbol{a}} W_{\boldsymbol{k},\boldsymbol{a}}, \\ a_j = Q_{\boldsymbol{c}} K_j^{\boldsymbol{c}\top} + Q_{\boldsymbol{c}} K_{\boldsymbol{s}}^{\top} + K_j^{\boldsymbol{c}} Q_{\boldsymbol{s}} + Q_{\boldsymbol{c}} K_{\boldsymbol{a}}^{\top} + K_j^{\boldsymbol{c}} Q_{\boldsymbol{a}} \\ \mathrm{head}_i = \mathrm{softmax}\left(\frac{\boldsymbol{a}}{\sqrt{5 d_H}}\right) H W_{\boldsymbol{v},\boldsymbol{c}}, \\ \mathbf{z}_{\boldsymbol{c}} = [\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h] W_O, \end{gathered} \quad (7)$$

where $\mathbf{u} = [\mathbf{u}_s, \mathbf{u}_a]$, a_j is the j-th element of $\boldsymbol{a}$, and $K_j^{\boldsymbol{c}}$ represents the j-th row of $K_{\boldsymbol{c}}$. The resulting $\mathbf{z}_c$ 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 $\boldsymbol{z}_s$ and the aspect vectors $\boldsymbol{z}_a$, 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 $\{\boldsymbol{x}^A, \boldsymbol{x}^B\}_{A,B \in G_k, A \neq B}$. 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:

$$\mathbf{z}^A = \mathrm{enc}(\boldsymbol{x}^A), \quad \mathbf{z}^B = \mathrm{enc}(\boldsymbol{x}^B). \quad (8)$$

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

[2] `https://pytorch.org/docs/stable/generated/torch.nn.GaussianNLLLoss.html`

independent components that encode the desiderata. If $\mathbf{z}_s$ and $\mathbf{z}_a$ are parameterized independent components triggered by $\mathbf{u}_s$ and $\mathbf{u}_a$ respectively, the substitution of $\mathbf{u}_a^B$ with $\mathbf{u}_a^A$ can be regarded as soft exchanges between $\mathbf{z}_a^A$ and $\mathbf{z}_a^B$.

We thus substitute $\mathbf{u}_a^B$ with $\mathbf{u}_a^A$ to cause changes in $\mathbf{z}_c^B$. This substitution will also be reflected by changes in $\mathbf{z}_a^B$. 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 balance the number of training examples unless it is content text generated from the swapped bottleneck representation.

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

$$
\begin{aligned}
Q_{\boldsymbol{s}}^B &= \mathbf{u}_{\boldsymbol{s}}^B W_{\boldsymbol{q},\boldsymbol{s}}, \quad K_{\boldsymbol{s}}^B = \mathbf{u}_{\boldsymbol{s}}^B W_{\boldsymbol{k},\boldsymbol{s}}, \\
Q_a^A &= \mathbf{u}_a^A W_{\boldsymbol{q},\boldsymbol{a}}, \quad K_a^A = \mathbf{u}_a^A W_{\boldsymbol{k},\boldsymbol{a}}, \\
a_j &= Q_{\boldsymbol{c}} K_j^{\boldsymbol{c}\top} + Q_{\boldsymbol{c}} K_{\boldsymbol{s}}^{B\top} + K_j^{\boldsymbol{c}} Q_s^B + Q_{\boldsymbol{c}} K_a^{A\top} + K_j^{\boldsymbol{c}} Q_a^A, \\
\text{head}_i &= \text{softmax}\left(\frac{\boldsymbol{a}}{\sqrt{5d_H}}\right) H W_{\boldsymbol{v},\boldsymbol{c}}, \\
\mathbf{z}_c^B &= [\text{head}_1, \text{head}_2, \ldots, \text{head}_h] W_O, \\
\mathbf{z}_s^B &= \text{softmax}\left(\frac{\mathbf{u}_s^B W_{\boldsymbol{q},\boldsymbol{s}} (K_{\mathsf{CLS}})^\top}{\sqrt{d_H}}\right) V_{\mathsf{CLS}}, \\
\mathbf{z}_a^B &= \text{softmax}\left(\frac{Q_a^A (K_a^A)^\top}{\sqrt{d_H}}\right) \mathbf{u}_a^A W_{\boldsymbol{v},\boldsymbol{a}},
\end{aligned}
$$

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

$$
\mathcal{L} = \mathcal{L}_c^A + \lambda_s \mathcal{L}_s^A + \lambda_a \mathcal{L}_a^A + \lambda_B \mathcal{L}_c^B, \qquad (9)
$$

where λ_s, λ_a and λ_B are hyper-parameters controlling the importance of each desirable property. In our experiments, we choose $\lambda_s = \lambda_a = 1$ 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.

Aspect Group	**Pro-Vax**	**Anti-Vax**	**Neutral**
	CMF		
Care/Harm	70	11	2
Fairness/Cheating	25	18	13
Loyalty/Betrayal	25	0	5
Authority/Subversion	20	46	13
Purity/Degradation	2	15	0
Liberty/Oppression	6	62	5
Non-moral	167	47	41
	VAD		
Health Institution	400	84	36
Personal Experience	381	16	3
Vaccines Save Lives	12	1	0
(Adverse) Side Effects	179	256	63
Immunity Level	433	113	52
Economic Effects	23	12	5
Personal Freedom	5	18	7
Moral Attitudes	5	43	2

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

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

[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 $\mathrm{NMI} = \big(2 \times \mathrm{I}(y;\hat{y})\big) / \big(\mathrm{H}(y) + \mathrm{H}(\hat{y})\big)$, where $\mathrm{I}(y;\hat{y})$ denotes the mutual information between the ground-truth labels and the predicted labels, $\mathrm{H}(\cdot)$ 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

$$\mathrm{AvgBS} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\binom{|G_k|}{2}}\sum_{\substack{i,j\in G_k\\ i<j}} \mathrm{BS}(\mathrm{tweet}_i, \mathrm{tweet}_j), \quad (10)$$

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 $K = 8$.

Models	CMF			VAD		
	CA	NMI	Avg BS	CA	NMI	Avg BS
SBERT-k-means	49.2	47.6	18.2	60.5	58.3	19.2
SBERT-k-medoids	50.8	48.1	18.5	62.1	60.1	19.5
SBERT-AHC	51.7	48.5	18.9	64.4	61.2	20.9
AutoBot-k-means	49.2	47.4	18.5	62.8	60.4	20.1
AutoBot-k-medoids	52.5	49.5	19.5	65.6	62.5	20.7
AutoBot-AHC	52.5	48.5	18.9	63.5	60.8	20.5
DS-C-k-means	50.0	47.7	18.5	63.5	60.5	20.7
DS-C-k-medoids	52.5	48.3	18.8	64.7	61.9	21.3
DS-C-k-AHC	50.8	47.8	18.6	64.4	61.5	21.7
VADet	51.7	47.9	18.0	65.4	61.4	20.7
SCCL	48.3	46.9	18.2	63.2	60.8	19.9
RoBERTa-k-means	35.0	35.2	15.0	45.8	46.6	15.7
DeBERTa-k-means	35.8	37.1	15.2	47.7	47.4	16.2
DOC-k-means	51.7	47.8	18.5	64.2	60.7	20.3
DOC-k-medoids	**54.2**	**51.0**	**20.7**	**66.7**	63.1	21.4
DOC-AHC	52.5	49.1	19.1	**66.7**	**63.6**	**22.8**

Table 2: Clustering results. Representation learning models are listed with the affiliated clustering methods.

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

Models	VAD → CMF			CMF → VAD		
	CA	NMI	Avg BS	CA	NMI	Avg BS
SBERT-AHC	51.6	49.8	19.3	52.4	50.5	17.9
AutoBot-k-medoids	53.1	50.6	20.1	53.7	51.0	18.1
DS-C-k-medoids	54.1	51.2	20.2	54.9	52.4	19.0
VADet	53.5	50.1	19.6	55.2	52.8	19.3
SCCL	48.6	47.0	18.5	53.6	51.6	18.5
DOC-k-medoids	**55.3**	**51.9**	**21.7**	**56.2**	**53.8**	**19.5**
DOC-AHC	53.5	50.4	19.8	55.8	53.7	19.2

Table 3: Cross-dataset evaluation results. Each representation learning model is listed with the most performant clustering method.

Models	CMF		VAD	
	Micro F1	Macro F1	Micro F1	Macro F1
RoBERTa	72.3±.5	71.2±.4	76.7±.1	75.9±.1
DeBERTa	**74.0**±.6	**73.5**±.6	77.8±.2	**76.8**±.2
DOC-AHC	73.5±.6	72.7±.6	**78.0**±.2	**76.8**±.2

Table 4: Stance classification results.

clusters as 7 and 8, respectively. The alignment between the clustered groups and gold labels is solved by the Hungarian algorithm. Note that direct aspect classification across datasets would not be possible since an accurate mapping between the two sets of classes cannot be established. Table 3 reports the performance of cross-dataset clustering. Our metrics of interest are still CA, NMI and averaged BERTScore. All the methods show a performance drop on VAD overall, while the performance on CMF turns out to be a bit higher. DOC-k-medoids achieved competitive results across the datasets, demonstrating that clustering-friendly representations disentangle the opinions and, as a result, can integrate unknown aspects.

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 $\boldsymbol{z}_s$, 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.

Model	CMF		VAD	
	CA	AvgBS	CA	AvgBS
Component				
DOC-k-means	**51.7**	**18.5**	**64.2**	**20.3**
w/o pre-trained LM	43.3	16.2	48.4	16.7
w/o inductive bias	50.0	18.0	62.3	19.2
w/o swapped codes	50.8	17.8	62.8	19.0
Choice of Context Vectors				
`MLP`	**51.7**	**18.5**	**64.2**	**20.3**
`CLS`	50.0	17.6	63.2	19.5
`MEAN`	48.3	17.4	60.7	18.7

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

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., $\boldsymbol{u}_s$ and $\boldsymbol{u}_a$. 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.

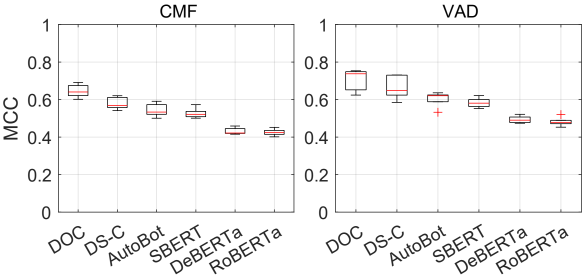


Figure 2: Boxplots of MCC for all representation learning models, over the 5 runs. The representations are used for k-means clustering in the Euclidean space. A high MCC score indicates the strong correlation between $dist(z_a, \bar{z}_a^k)$ and $z_a \in G_k$.

Evaluation of Disentangled Representations As with the nonlinear ICA community (Khemakhem et al., 2020), we use Mean Correlation Coefficient (MCC) to quantify the extent to which DOC managed to learn disentangled representations. Here, the Point-Biserial Correlation Coefficient between $dist(z_a, \bar{z}_a^k)$ (i.e., the distance between 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 $k \in \{1, 2, \dots, K\}$, 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

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Iman Munire Bilal, Bo Wang, Maria Liakata, Rob Procter, and Adam Tsakalidis. 2021. Evaluation of thematic coherence in microblogs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6800–6814, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Ranganathan Chandrasekaran, Rashi Desai, Harsh Shah, Vivek Kumar, and Evangelos Moustakas. 2022. Examining public sentiments and attitudes toward covid-19 vaccination: Infoveillance study using twitter posts. *JMIR Infodemiology*, 2(1):e33909.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Koustava Goswami, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. Unsupervised deep language and dialect identification for short texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1606–1617, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Daniella Horan, Eitan Richardson, and Yair Weiss. 2021. When is unsupervised disentanglement possible? In *Advances in Neural Information Processing Systems*, volume 34, pages 5150–5161. Curran Associates, Inc.

Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. 2020. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6989–6999, Online. Association for Computational Linguistics.

Amir Hussain, Ahsen Tahir, Zain Hussain, Zakariya Sheikh, Mandar Gogate, Kia Dashtipour, Azhar Ali, and Aziz Sheikh. 2021. Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study. *J Med Internet Res*, 23(4):e26627.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR.

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3581–3589, Cambridge, MA, USA. MIT Press.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Florian Kunneman, Mattijs Lambooij, Albert Wong, Antal van den Bosch, and Liesbeth Mollema. 2020. Monitoring stance towards vaccination in twitter messages. *BMC medical informatics and decision making*, 20(1):1–14.

Kaufman Leonard and J Rousseeuw Peter. 1990. Finding groups in data: an introduction to cluster analysis. *Probability and Mathematical Statistics. Applied Probability and Statistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR.

Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. 2020a. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR.

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. 2020b. Disentangling factors of variations using few labels. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Junru Lu, Xingwei Tan, Gabriele Pergola, Lin Gui, and Yulan He. 2022. Event-centric question answering via contrastive learning and invertible event transformation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2377–2389, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*

Neural Information Processing Systems, volume 29. Curran Associates, Inc.

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.

Sebastião Miranda, Artūrs Znotiņš, Shay B. Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4535–4544, Brussels, Belgium. Association for Computational Linguistics.

Ivan Montero, Nikolaos Pappas, and Noah A. Smith. 2021. Sentence bottleneck autoencoders from transformer language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1822–1831, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, Seattle, United States. Association for Computational Linguistics.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. 2020. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems*, volume 33, pages 7198–7211. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Uri Shaham, Kelly P. Stanton, Henry Li, Ronen Basri, Boaz Nadler, and Yuval Kluger. 2018. Spectralnet: Spectral clustering using deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online. Association for Computational Linguistics.

Prateek Sircar, Aniket Chakrabarti, Deepak Gupta, and Anirban Majumdar. 2022. Distantly supervised aspect clustering and naming for E-commerce reviews. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 94–102, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Yaling Tao, Kentaro Takagi, and Kouta Nakata. 2021. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

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.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA. PMLR.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021a. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430, Online. Association for Computational Linguistics.

Haidong Zhang, Wancheng Ni, Meijing Zhao, and Ziqi Lin. 2019. Cluster-gated convolutional neural network for short text classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1002–1011, Hong Kong, China. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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*.

Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. 2021d. Disentangling representations of text by masking transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 778–791, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Runcong Zhao, Miguel Arana-catania, Lixing Zhu, Elena Kochkina, Lin Gui, Arkaitz Zubiaga, Rob Procter, Maria Liakata, and Yulan He. 2023. PANACEA: An automated misinformation detection system on COVID-19. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Lixing Zhu, Zheng Fang, Gabriele Pergola, Robert Procter, and Yulan He. 2022. Disentangled learning of stance and aspect topics for vaccine attitude detection in social media. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1580, Seattle, United States. Association for Computational Linguistics.

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 $d_{\mathbf{z}} = 1024$. 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 $\boldsymbol{z}_a$ or the content vectors $\boldsymbol{z}$ (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

CMF		
Tweet	**Argumentative Pattern**	**Aspect Group**
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
VAD		
Tweet	**Aspect Span**	**Aspect Group**
Study reports Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19.	Oxford/AstraZeneca vaccine is protective against Brazilian P1 strain of COVID19	Immunity Level
@user @user @user team, told Reuters while the government admits, it is unknown whether COVID19 mRNA Vaccine BNT162b2 has an impact on fertility.	COVID19 mRNA Vaccine BNT162b2 has an impact on fertility	(Adverse) Side Effects

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 subsequence of the annotated tweets.

Latent Vector	CMF		VAD	
	CA	AvgBS	CA	AvgBS
DOC-k-means-z_a	51.7	18.5	64.2	20.3
DOC-k-means-z	48.3	17.5	60.7	18.7
DOC-k-medoids-z_a	**54.2**	**20.7**	**66.7**	21.4
DOC-k-medoids-z	50.8	18.0	61.4	18.9
DOC-AHC-z_a	52.5	19.1	**66.7**	**22.8**
DOC-AHC-z	49.2	17.8	61.9	19.0

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

Qualitative Results We illustrate in Figure A1 and Figure A2 the clustering results and the latent space of the entangled/disentangled representation projected by the t-SNE method. Figure A1(a-b) display the cluster assignments after permutation, whereas Figure A2(a-b) show the ground-truth labels. The class labels are rendered by colours whose detailed mapping is provided in Figure A2. From Figure A1, we see clear improvements in terms of clustering quality on both datasets when the model is compared against the DeBERTa-averaged-embedding. Figure 2 shows more separated groups thanks to the disentangled representation, providing strong distance-based discrimination for the clustering algorithms. As a result, simple clustering methods like k-means can achieve competitive results against deep clustering methods (i.e., SCCL and VAD), which have access to weak labels or data augmentations.

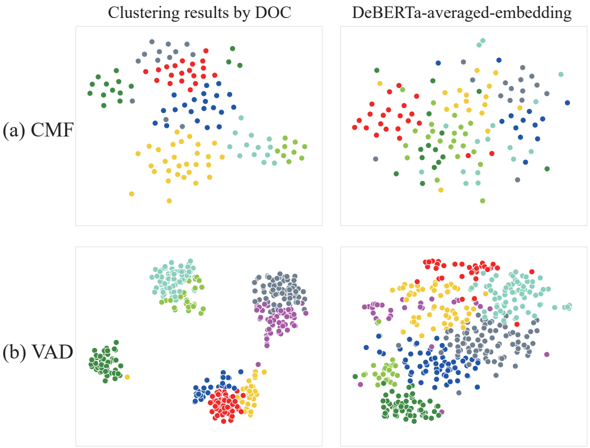


Figure A1: 2-D plots of the data points projected by t-SNE.

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.

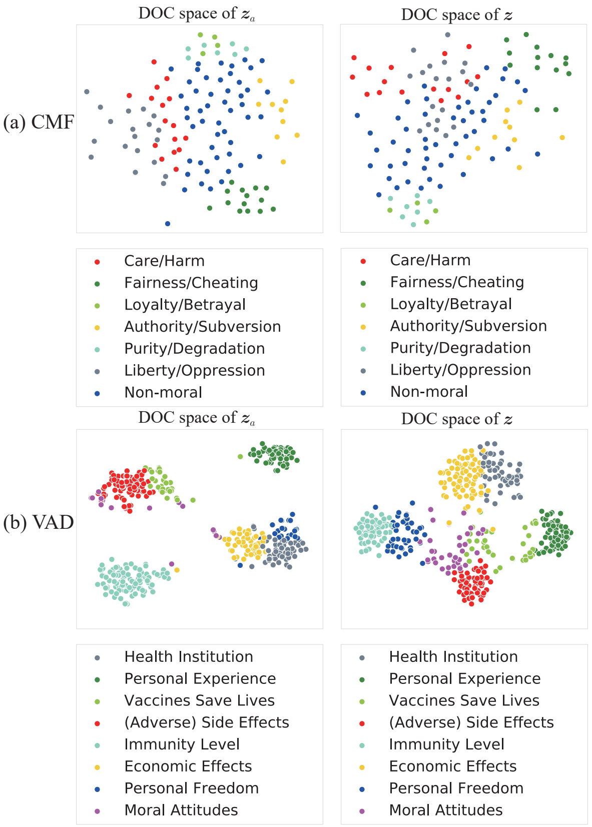


Figure A2: t-SNE plots on CMF and VAD. Each dot is a tweet encoded using either the disentangled aspect vector z_a (left subfigure) or the latent content vector z (right subfigure). Different colors indicate the true aspect category labels.

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?

Section 3

☑ 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?

Section 4

☑ 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.