Supplementary Materials

A Proof of the Two Bounds for the Information Bottleneck

A.1 The Lower Bound for $I(Z;Y)$

$$
\begin{aligned}
I(Z;Y) &= \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)} \\
&= \sum_{y,z} p(y,z) \log \frac{p(y \mid z)}{p(y)} \\
&= \sum_{y,z} p(y,z) \log p(y \mid z) - \sum_{y,z} p(y,z) \log p(y)
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
&\sum_{y,z} p(y,z) \log p(y \mid z) \\
=&\sum_{y,z} p(y,z) \log \frac{p(y \mid z) q_\phi(y \mid z)}{q_\phi(y \mid z)} \\
=&\sum_{y,z} p(y,z) \log q_\phi(y \mid z) + \sum_{z} p(z) \mathrm{KL}[p(y \mid z) \| q_\phi(y \mid z)] \\
\geqslant&\sum_{y,z} p(y,z) \log q_\phi(y \mid z)
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
I(Z;Y) &\geqslant \sum_{y,z} p(y,z) \log q_\phi(y \mid z) - \sum_{y,z} p(y,z) \log p(y) \\
&= \sum_{y,z} p(y,z) \log \frac{q_\phi(y \mid z)}{p(y)} \\
&= \sum_{y,z,r} p(r,y,z) \log \frac{q_\phi(y \mid z)}{p(y)} \\
&= \sum_{y,z,r} p(r,y) p(z \mid r) \log \frac{q_\phi(y \mid z)}{p(y)}
\end{aligned} \tag{12}
$$

A.2 The Upper Bound for $I(Z;R)$

$$
\begin{aligned}
I(Z;R) &= \sum_{r,z} p(r,z) \log \frac{p(r,z)}{p(r)p(z)} \\
&= \sum_{r,z} p(r,z) \log \frac{p(z \mid r)}{p(z)} \\
&= \sum_{r,z} p(r,z) \log p(z \mid r) - \sum_{r,z} p(r,z) \log p(z)
\end{aligned} \tag{13}
$$

By replacing $p(z)$ with a prior distribution of z, $r_\psi(z)$, we have

$$
\sum_{r,z} p(r,z) \log p(z) \geqslant \sum_{r,z} p(x,z) \log r_\psi(z) \tag{14}
$$

Then we can obtain an upper bound of $I(Z;R)$,

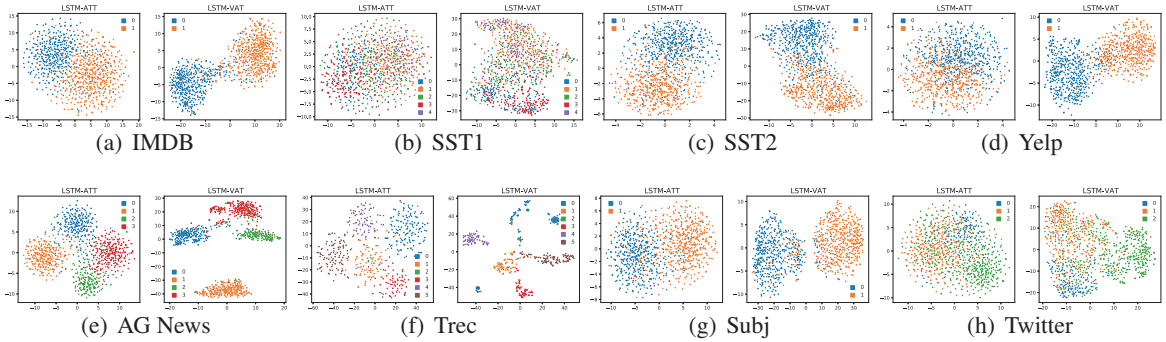
$$
\begin{aligned}
I(Z;R) \leqslant& \sum_{r,z} p(r,z) \log p(z \mid r) - \sum_{r,z} p(r,z) \log r_\psi(z) \\
=& \sum_{r} p(r) \mathrm{KL}[p(z \mid r) \| r_\psi(z)] \\
=& \mathbb{E}_{p(r)}[\mathrm{KL}[p(z \mid r) \| r_\psi(z)]]
\end{aligned} \tag{15}
$$

Dataset	0	1	2	3	4	5
Twitter	70.75	71.62	70.96	70.67	71.06	70.98
AG News	91.88	91.75	92.04	91.97	91.69	91.78
SST-1	46.29	46.61	47.42	47.10	46.92	46.88
SST-2	84.73	84.62	85.61	86.22	86.11	85.94
Subj	90.80	91.10	90.80	90.30	90.70	90.40
Trec	91.00	91.20	91.60	92.00	92.40	92.80
IMDB	88.16	88.98	88.22	88.84	88.14	88.60
Yelp	95.06	95.32	95.12	95.04	94.99	94.57

Table 7: The accuracy of our LSTM-VAT model with different iteration number.

Dataset	0	1	2	3	4	5
Twitter	74.84	75.26	77.71	77.13	76.68	76.76
AG News	93.41	93.50	93.43	93.71	93.20	93.34
SST-1	51.13	51.58	51.86	51.99	51.54	51.22
SST-2	91.16	91.21	91.43	91.93	91.98	91.76
Subj	96.20	96.20	96.10	96.70	96.40	96.30
Trec	96.40	96.80	97.20	96.80	96.80	96.40
IMDB	91.81	92.06	92.11	92.09	91.92	91.96
Yelp	97.20	97.35	97.36	97.28	97.32	97.27

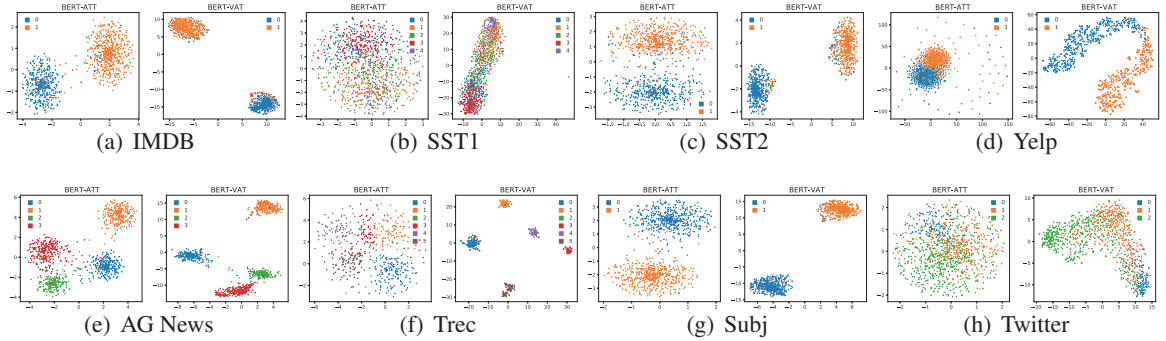
Table 8: The accuracy of our BERT-VAT model with different iteration number.



(a) IMDB (b) SST1 (c) SST2 (d) Yelp

(e) AG News (f) Trec (g) Subj (h) Twitter

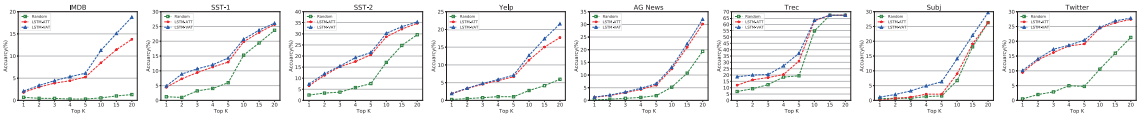
Figure 9: Visualization of sentence representation obtained from LSTM-ATT and LSTM-VAT. We use t-SNE to transfer 100-dimensional feature space into two-dimensional space.



(a) IMDB (b) SST1 (c) SST2 (d) Yelp

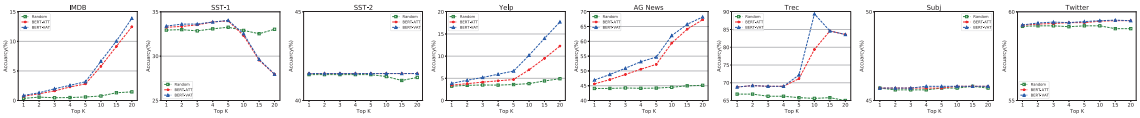
(e) AG News (f) Trec (g) Subj (h) Twitter

Figure 10: Visualization of sentence representation obtained from BERT-ATT and BERT-VAT. We use t-SNE to transfer 768-dimensional feature space into two-dimensional space.



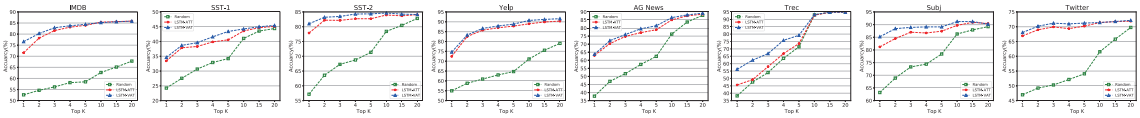
(a) IMDB (b) SST1 (c) SST2 (d) Yelp (e) AG News (f) Trec (g) Subj (h) Twitter

Figure 11: The influence of Top-K for LSTM-based models in terms of AOPC.



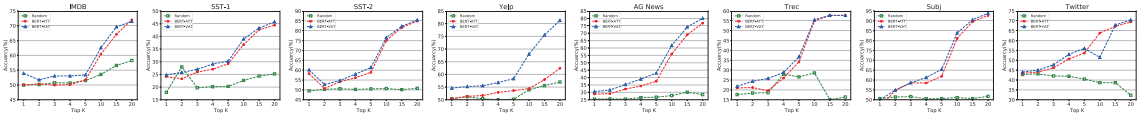
(a) IMDB (b) SST1 (c) SST2 (d) Yelp (e) AG News (f) Trec (g) Subj (h) Twitter

Figure 12: The influence of Top-K for BERT-based models in terms of AOPC.



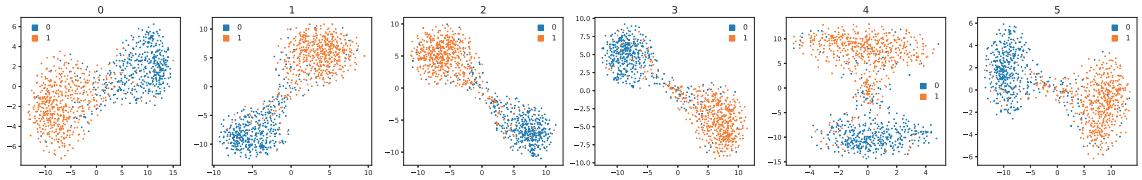
(a) IMDB (b) SST1 (c) SST2 (d) Yelp (e) AG News (f) Trec (g) Subj (h) Twitter

Figure 13: The influence of Top-K for LSTM-based models in terms of Post-hoc.

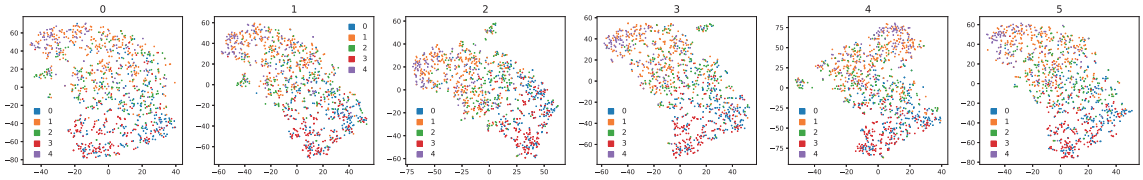


(a) IMDB (b) SST1 (c) SST2 (d) Yelp (e) AG News (f) Trec (g) Subj (h) Twitter

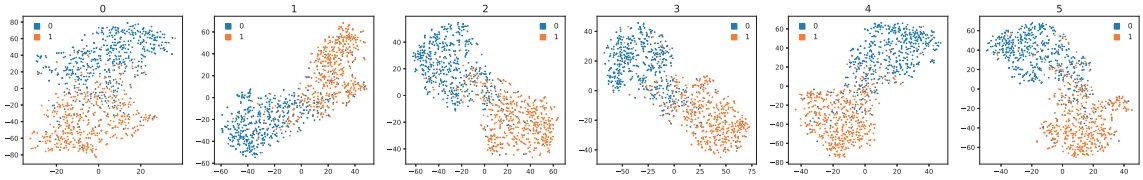
Figure 14: The influence of Top-K for BERT-based models in terms of Post-hoc.



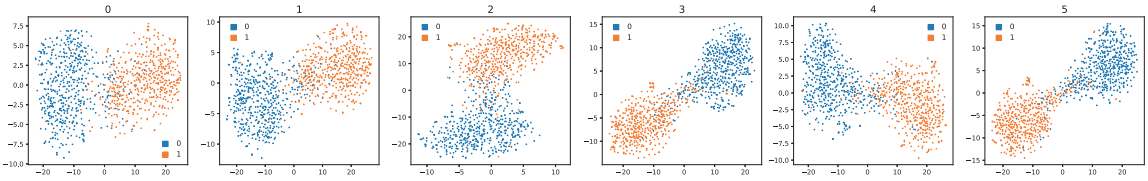
(a) IMDB



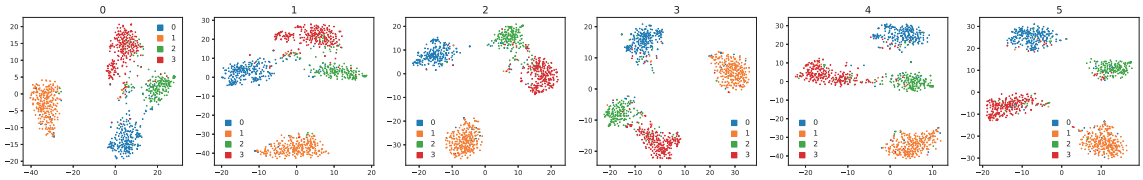
(b) SST1



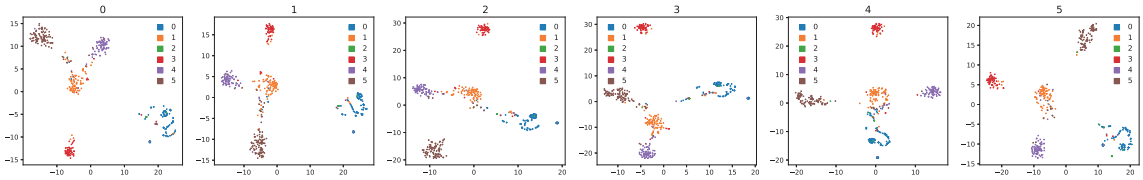
(c) SST2



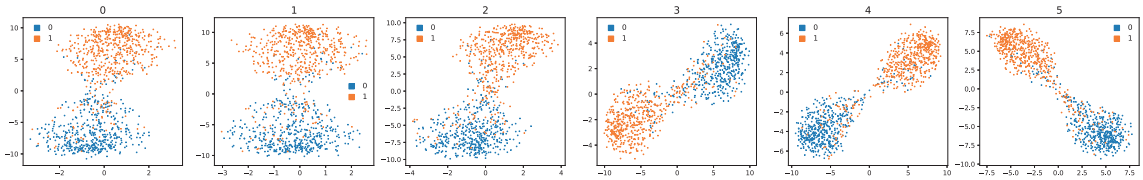
(d) Yelp



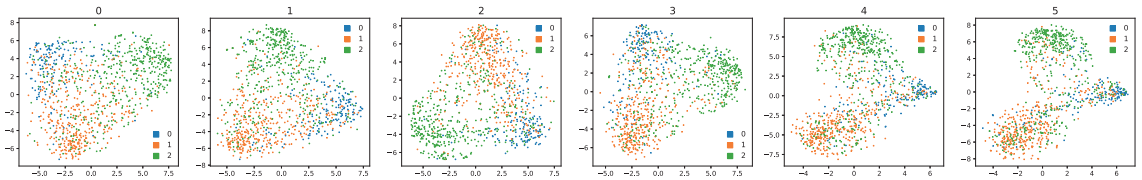
(e) AG News



(f) Trec

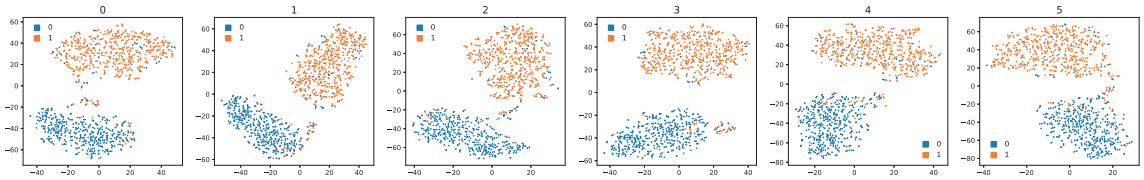


(g) Subj

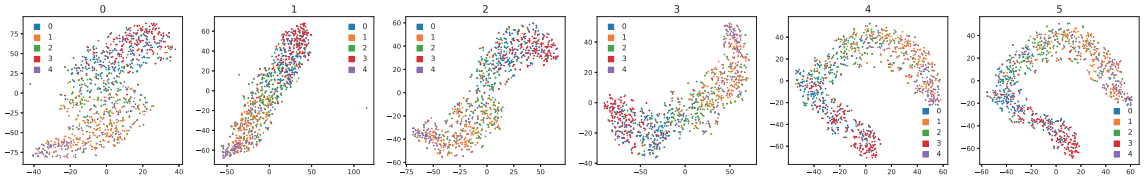


(h) Twitter

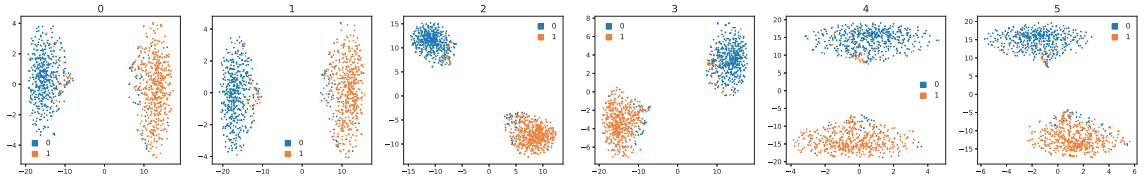
Figure 15: Visualization of sentence representation obtained from LSTM-VAT with different iterations. We use t-SNE to transfer 100-dimensional feature space into two-dimensional space.



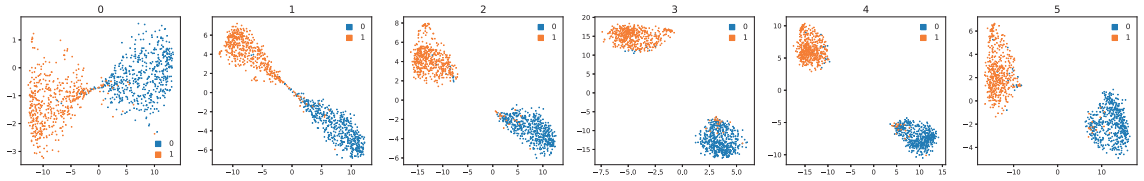
(a) IMDB



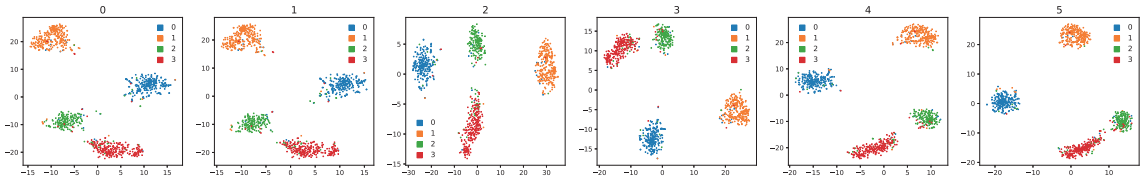
(b) SST1



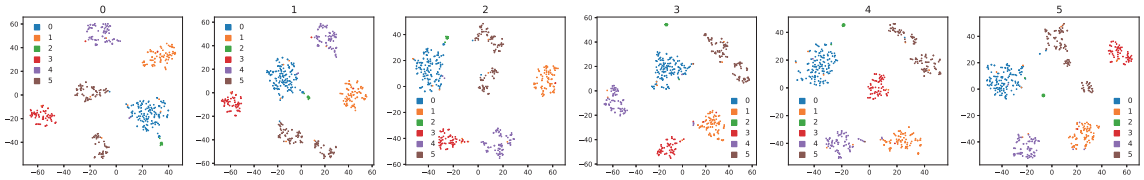
(c) SST2



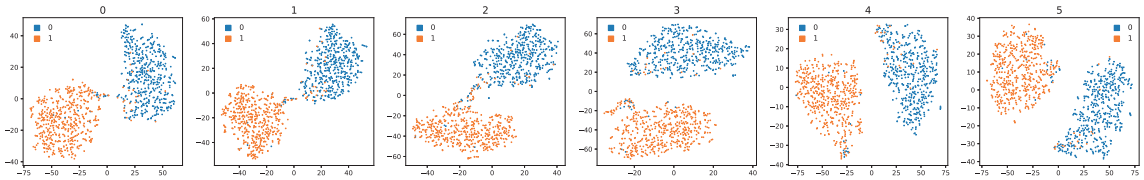
(d) Yelp



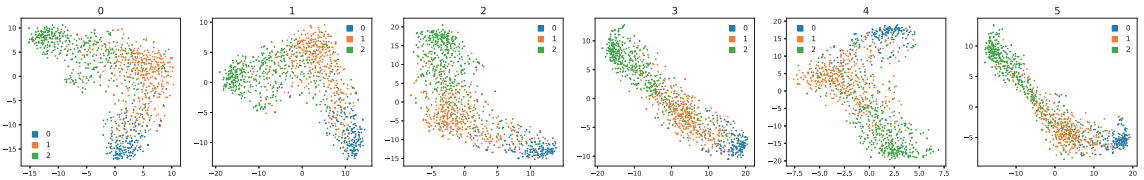
(e) AG News



(f) Trec



(g) Subj



(h) Twitter

Figure 16: Visualization of sentence representation obtained from BERT-VAT with different iterations. We use t-SNE to transfer 768-dimensional feature space into two-dimensional space.