A. Algorithms

A.1. Pseudocode

We present pseudocode for each of the algorithms included in our experiments.

Algorithm: SynTF Term-Frequency Vector Synthesis (Weggenmann and Kerschbaum, 2018)

Input: Document vector θ_t, desired output length n, privacy parameter $\epsilon > 0$, vocabulary V, rating function $\rho : V \times V \to [0, 1]$.
Output: Synthetic tf vector $s \in \mathbb{N}^{|V|}$ with $|s| = n$
1: **for** $i \in \{1, \ldots, n\}$ **do**
2:      Randomly sample word $v_i = R\, \mathsf{Cat}(\theta_t)$
3:      Randomly choose synonym $w_i = R\, \mathsf{E}_\epsilon, \rho(v_i)$
4: **end for**
5: $s_w = |i \in [1, n] : w_i = w|$ for all $w \in V$ ▷ count synonyms
6: **return** s

Algorithm: Calibrated Multivariate Perturbation Mechanism (Feyisetan et al., 2020)

Input: String $x = w_1 w_2 \ldots w_n$, privacy parameter $\epsilon > 0$, word set $\mathcal{W}$.
Output: Privatized string $\tilde{s}$
1: **for** $i \in \{1, \ldots, n\}$ **do**
2:      Compute embedding $\varphi_i = \varphi(w_i)$
3:      Perturb embedding to obtain $\hat{\varphi}_i = \varphi_i + \mathcal{N}$ with noise density $p_\mathcal{N}(z) \propto \exp(-\epsilon\|z\|)$
4:      Obtain perturbed word $\hat{w}_i = \arg\min_{u \in \mathcal{W}} |\varphi(u) - \hat{\varphi}_i|$
5:      Insert $\hat{w}_i$ in i^{th} position of $\hat{x}$
6: **end for**
7: **return** $\hat{x}$

Algorithm: The Mahalanobis Mechanism (Xu et al., 2020)

Input: String $s = w_1 w_2 \ldots w_n$, privacy parameter $\epsilon > 0$, scaled sample covariance matrix Σ, tuning parameter $\lambda \in [0, 1]$, word set $\mathcal{W}$.
Output: Privatized string $\tilde{s}$
1: **for** $i \in \{1, \ldots, n\}$ **do**
2:      Sample Z from $f_Z(z) \propto \exp(-\epsilon\|z\|_{RM})$
3:      Obtain the perturbed embedding $\hat{\phi}_i = \phi(w_i) + Z$.
4:      Replace w_i with $\hat{w}_i = \arg\min_{w \in \mathcal{W}} \|\phi(w) - \hat{\phi}_i\|_2$.
5: **end for**
6: **return** $\tilde{s} = \hat{w}_1 \hat{w}_2 \ldots \hat{w}_n$

Algorithm: SanText Base Mechanism (Yue et al., 2021)

Input: A private document $D = \langle x_i \rangle_{i=1}^L$ and a privacy parameter $\epsilon > 0$.
Output: Sanitized document $\hat{D}$
1: Derive token vectors $\phi(x_i)$ for $i \in [1, L]$;
2: **for** $i = 1, \ldots, L$ **do**
3:      Run $M(x_i)$ to sample sanitized token y_i with $\Pr[M(x) = y] = C_x \cdot e^{-\frac{1}{2}\epsilon \cdot d_{\mathsf{euc}}(\phi(x), \phi(y))}$
4: **end for**
5: Output sanitized $\hat{D}$ as $\langle y_i \rangle_{i=1}^L$

Algorithm: The Truncated Gumbel Mechanism (Xu et al., 2021a)

Input: String $s = w_1 w_2 \ldots w_n$, privacy parameter $\epsilon > 0$, word set $\mathcal{W}$.
Output: Privatized string $\tilde{s}$
  1: $\Delta = \max_{w,w' \in \mathcal{W}} \|\phi(w) - \phi(w')\|$, $\Delta_0 = \min_{w,w' \in \mathcal{W}} \|\phi(w) - \phi(w')\|$ ▷ max, min word distance
  2: b $= \frac{2\Delta}{\min\{W(2\alpha\Delta), \log_e(\alpha\Delta_0)\}}$, with $\alpha = \frac{1}{3}\left(\epsilon - \frac{2(1+\log\|\mathcal{W}\|)}{\Delta_0}\right)$ and W = principal branch of Lambert W function.
  3: **for** $i \in \{1, \ldots, n\}$ **do**
  4:      $k \sim \text{TruncatedPoisson}(\log |\mathcal{W}|; 1, |\mathcal{W}|)$ and find k closest words to w_i as $\mathbf{u}$ with $u_1 = w_i$
  5:      Compute k distances $\mathbf{d}$, where $d_j = \|w_i - u_j\|_2$
  6:      $\hat{w}_i = u_j$, where $j = \underset{\{d_1+g_1,\ldots,d_k+g_k\}}{\text{argmin}}$ and $g_1, \ldots, g_k \sim_{i.i.d.} \text{TruncatedGumbel}(0, b, \Delta)$
  7: **end for**
  8: **return** $\tilde{s} = \hat{w}_1 \hat{w}_2 \ldots \hat{w}_n$.

Algorithm: The Vickrey Mechanism (k=2) (Xu et al., 2021b)

Input: String $s = w_1 w_2 \ldots w_n$, metric d, privacy parameter $\epsilon > 0$, tuning parameter $t \in [0,1]$, word set $\mathcal{W}$.
Output: Privatized string $\tilde{s}$
  1: **for** $i \in \{1, \ldots, n\}$ **do**
  2:      Sample Z from $p(z) \propto \exp\{-\epsilon d(z, 0)\}$
  3:      Obtain the perturbed embedding $\hat{\phi}_i = \phi(w_i) + Z$
  4:      Let $\tilde{w}_{i1} = \underset{w \in \mathcal{W}\setminus\{w_i\}}{\text{argmin}} \|\hat{\phi}_i - \phi(w)\|_2$ and $\tilde{w}_{i2} = \underset{w \in \mathcal{W}\setminus\{w_i, \tilde{w}_{i1}\}}{\text{argmin}} \|\hat{\phi}_i - \phi(w)\|_2$
  5:

$$\text{Set } \hat{w}_i = \begin{cases} \tilde{w}_{i1}, & \text{w/ Pr } p(t, \hat{\phi}_i) \\ \tilde{w}_{i2}, & \text{w/ Pr } 1 - p(t, \hat{\phi}_i) \end{cases}, \text{where } p(t, \hat{\phi}_i) = \frac{(1-t)\|\phi(\tilde{w}_{i2}) - \hat{\phi}_i\|_2}{t\|\phi(\tilde{w}_{i1}) - \hat{\phi}_i\|_2 + (1-t)\|\phi(\tilde{w}_{i2}) - \hat{\phi}_i\|_2}$$

  6: **end for**
  7: **return** $\tilde{s} = \hat{w}_1 \hat{w}_2 \ldots \hat{w}_n$.

Algorithm: TEM: Metric Truncated Exponential Mechanism (Carvalho et al., 2023)

Input: Word set $\mathcal{W}$, input word $w \in W$, truncation threshold γ, metric $d_W : W \times W \to \mathbb{R}^+$, and privacy parameter $\epsilon > 0$.
Output: Privatized string $\tilde{s}$
  1: Given input w, obtain the set L_w such that each word $w_i \in L_w$ satisfies $d_W(w, w_i) \leq \gamma$
  2: Set the score $f(w, w_i)$ of each $w_i \in L_w$ as $f(w, w_i) = -d_W(w, w_i)$
  3: Create a $\perp$ element with score $f(w, \perp) = -\gamma + 2\ln(|W \setminus L_w|)/\epsilon$
  4: **for** each word $w_i \in L(x) \cup \perp$ **do**
  5:      add Gumbel noise with mean 0 and scale $2/\epsilon$ to score $f(w, w_i)$
  6: **end for**
  7: Select $\hat{w}$ as the element with maximum noisy score from $L(x) \cup \perp$
  8: **if** $\hat{w} = \perp$ **then**
  9:      **return** random sample of $W \setminus L_w$
10: **else**
11:      **return** $\hat{w}$
12: **end if**

A.2. Algorithm-specific Parameters

In our experiments, we used the following algorithm-specific parameters (beyond ϵ):
SynTF: synonyms taken from NLTK WordNet
Mahalanobis: $\lambda = 0.2$
Vickrey: $t = 0.5$
TEM: $\gamma = 0.5$

B. Privacy Metric Results

In Tables 5 and 6, we present the complete set of privacy metrics as evaluated in our study. For each combination of (*task, dimension, epsilon*), we **bold** the best score, which is governed by whether a lower (↓) or a higher (↑) score is better.

Task:	Sentiment Analysis (IMDb)														
Metric:	PD (N_w)↓	PD (S_w)↑	PP↑	CS↑	LOW↓	PD (N_w)↓	PD (S_w)↑	PP↑	CS↑	LOW↓	PD (N_w)↓	PD (S_w)↑	PP↑	CS↑	LOW↓
Epsilon:	1					5					10				
Dimension:						50									
SynTF	32.1	5.1	70.5	62.9	70.5	33.0	5.1	76.5	61.6	77.1	33.4	5.1	78.4	62.2	76.6
CMP	0.0	97.5	**98.2**	33.5	46.8	9.1	87.8	90.1	45.0	47.2	74.1	24.7	33.5	82.0	64.3
Mahalanobis	0.0	97.4	**98.2**	33.9	46.8	9.0	88.1	91.8	44.0	46.9	71.2	26.6	42.7	76.7	56.6
SanText	14.6	84.8	65.0	**70.7**	65.0	14.6	84.8	59.8	79.5	**13.3**	14.0	**85.4**	59.2	82.5	**29.1**
Gumbel	23.6	13.7	76.6	64.0	53.1	22.7	13.3	76.8	63.9	52.5	24.0	13.6	76.9	63.8	53.6
Vickrey	0.0	99.0	**98.2**	33.6	46.7	**3.5**	**91.8**	**95.0**	42.8	46.9	**5.8**	64.9	**91.0**	53.9	47.5
TEM	0.0	**99.9**	**98.2**	35.5	**2.6**	68.1	31.8	18.4	**87.3**	30.5	99.9	0.1	0.1	**99.0**	99.8
Dimension:						100									
SynTF	31.3	5.1	70.3	62.9	77.0	32.1	5.1	76.2	61.8	76.8	32.4	5.1	78.1	62.4	77.2
CMP	0.0	97.3	**98.2**	33.9	46.9	3.4	95.0	94.4	40.6	46.7	52.2	46.0	47.1	72.5	48.0
Mahalanobis	0.0	98.1	**98.2**	34.2	46.7	3.3	95.8	95.2	39.8	46.7	46.9	51.2	58.6	65.2	48.3
SanText	13.6	86.0	66.5	**70.6**	3.0	14.4	85.2	59.4	79.4	**9.9**	14.0	**85.2**	59.2	83.1	**26.8**
Gumbel	28.0	13.9	73.0	67.2	42.6	27.2	13.8	73.4	67.0	42.3	25.3	13.8	73.7	66.8	41.3
Vickrey	0.0	99.0	**98.2**	34.0	46.8	**2.0**	**96.7**	**96.5**	39.4	47.1	**6.4**	80.2	**90.5**	51.2	46.9
TEM	0.0	**99.9**	**98.2**	35.5	3.1	77.5	22.4	6.1	**95.1**	37.5	99.9	0.1	0.1	**99.0**	97.9
Dimension:						300									
SynTF	32.1	5.0	69.9	63.2	76.7	32.0	5.1	75.4	62.1	76.3	32.6	5.1	77.5	62.8	76.8
CMP	0.0	98.0	**98.2**	32.7	46.9	1.0	98.7	97.4	35.6	46.8	27.8	71.5	78.8	49.8	46.7
Mahalanobis	0.0	98.5	**98.2**	34.0	46.9	0.6	99.1	**97.7**	36.2	46.8	20.4	78.9	88.5	44.2	46.5
SanText	14.4	85.3	66.6	**70.6**	**2.7**	14.2	85.3	59.2	80.7	**16.6**	14.3	85.5	59.2	84.0	33.1
Gumbel	31.7	13.4	69.5	68.1	25.2	32.5	13.0	70.0	67.8	25.8	27.7	13.6	70.9	67.5	**25.0**
Vickrey	0.0	99.5	**98.2**	34.3	48.3	**0.5**	**99.2**	97.63	37.2	47.6	**5.8**	**91.0**	**92.7**	44.5	49.3
TEM	0.1	**99.8**	98.1	35.2	3.0	99.2	0.8	1.4	**98.7**	71.4	99.9	0.1	0.1	**99.9**	98.4

Table 5: Privacy scores for the Sentiment Anaylsis (IMDb) experiments.

Task:	Topic Classification (AG News)														
Metric:	PD (N_w)↓	PD (S_w)↑	PP↑	CS↑	LOW↓	PD (N_w)↓	PD (S_w)↑	PP↑	CS↑	LOW↓	PD (N_w)↓	PD (S_w)↑	PP↑	CS↑	LOW↓
Epsilon:	1					5					10				
Dimension:						50									
SynTF	55.1	4.6	60.7	**61.3**	61.7	54.8	4.7	66.4	58.6	60.3	55.2	4.8	68.4	58.1	60.2
CMP	0.0	97.4	**98.8**	20.3	16.0	12.4	83.9	85.9	37.9	17.2	75.8	21.3	23.1	84.5	52.5
Mahalanobis	0.0	97.7	**98.8**	20.9	16.5	10.9	85.3	88.4	35.5	16.2	75.0	22.2	30.3	79.3	36.8
SanText	3.9	95.0	47.2	60.0	5.7	**3.2**	**95.7**	26.5	81.0	27.5	**3.0**	**95.0**	25.7	84.3	25.7
Gumbel	25.6	13.2	73.0	57.7	27.2	26.9	13.0	73.2	57.6	29.3	26.0	12.7	73.4	57.6	28.2
Vickrey	0.0	98.8	**98.8**	20.8	16.1	4.7	90.4	**94.2**	33.2	**16.0**	5.1	57.6	**92.8**	44.0	**20.0**
TEM	0.1	**99.7**	**98.8**	21.2	**4.5**	87.2	12.6	7.5	**94.3**	69.3	100.0	0.0	0.5	**99.9**	99.3
Dimension:						100									
SynTF	55.4	4.5	60.6	**61.4**	61.6	55.2	4.6	66.1	58.8	59.8	55.4	4.6	68.1	58.3	59.9
CMP	0.0	97.9	98.8	21.1	17.3	5.7	93.0	92.3	31.2	16.7	60.2	37.2	35.3	75.0	25.1
Mahalanobis	0.0	98.3	98.8	21.5	17.7	5.5	93.0	93.7	29.6	**15.9**	58.8	38.9	46.5	66.4	**17.7**
SanText	3.9	95.4	47.2	59.8	5.4	4.2	94.8	25.8	81.6	20.8	**4.7**	**94.6**	25.6	85.0	56.4
Gumbel	29.8	13.0	70.7	60.8	22.8	28.7	12.5	71.0	60.6	21.5	27.6	12.6	71.5	60.3	22.3
Vickrey	0.0	98.9	**98.9**	21.3	16.6	**2.8**	**95.8**	**96.2**	28.8	16.9	6.8	78.1	**91.7**	41.5	**17.7**
TEM	0.1	**99.7**	98.8	20.8	**4.4**	93.6	6.4	2.1	**98.5**	81.4	99.8	0.2	0.2	**99.9**	98.3
Dimension:						300									
SynTF	55.7	4.6	60.2	**61.6**	61.2	55.7	4.6	65.3	59.1	61.7	55.4	4.7	67.7	58.5	60.2
CMP	0.0	98.1	**98.9**	19.1	16.5	2.2	97.0	97.3	23.8	16.0	34.9	63.7	69.2	48.3	15.6
Mahalanobis	0.0	98.2	**98.9**	20.8	17.5	1.5	98.1	97.9	23.9	16.4	26.8	71.8	83.5	36.9	15.0
SanText	3.6	95.5	47.2	59.8	6.3	4.0	95.5	25.8	82.0	32.8	**3.6**	95.7	25.7	85.4	57.7
Gumbel	32.0	12.6	68.8	60.5	23.9	31.2	12.8	69.2	60.2	23.5	30.7	13.2	69.9	59.6	23.9
Vickrey	0.0	98.8	**98.9**	19.6	16.8	**1.0**	**98.5**	**98.0**	23.5	**15.3**	6.6	**89.7**	**92.5**	33.6	**14.6**
TEM	0.4	**99.4**	98.6	20.5	**5.4**	99.9	0.1	0.51	**99.8**	98.0	99.8	0.2	0.13	**99.9**	98.4

Table 6: Privacy scores for the Topic Classification (AG News) experiments.

B.1. Percentage of English Words

As another metric to measure the effect of MLDP methods on word perturbation, we measured the percentage of English words existing in the perturbed datasets, as opposed to the baseline in the original ones. To calculate this metric, we leveraged the **words** corpus from *NLTK*. Note that a word not being counted as English does not necessarily mean that it is in a different language, but rather that it is not included as a standard word in our chosen corpus of English words.

As the true relation between privacy preservation via word perturbations and the resulting effect on English words is not well known and justifiable in comparison to our other metrics, we exclude this metric from our main analysis and instead include the results here. Nevertheless, it is interesting to observe that

some methods significantly reduce the number of English words (e.g., CMP and TEM). A more in-depth study into the relevance of these results can be a point of future investigation.

Dataset	IMDb			AG News		
Baseline	84.75			73.63		
Epsilon	1	5	10	1	5	10
SynTF	85.05	85.92	86.46	75.16	75.62	75.96
CMP	53.97	67.77	78.08	41.55	59.88	69.74
Mahalanobis	57.44	67.67	75.72	47.67	60.13	67.54
SanText	76.42	78.33	80.54	67.36	70.98	71.40
Gumbel	79.18	79.13	79.10	67.18	67.16	67.05
Vickrey	56.95	69.00	75.76	44.40	59.97	66.22
TEM	34.59	80.90	84.73	45.24	72.92	73.60

Table 7: Percentage of English words in perturbed datasets.

B.2. Privacy-Utility Composite Scores

The full set of Privacy-Utility Composite (PUC) scores is provided in Table 8.

Task:	Sentiment Analysis (IMDb)									Topic Classification (AG News)								
Tuning parameter:	$\alpha=0.75$			$\alpha=0.5$			$\alpha=0.25$			$\alpha=0.75$			$\alpha=0.5$			$\alpha=0.25$		
Epsilon:	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
Dimension: 50																		
SynTF	82.22	81.76	81.48	70.54	70.05	70.03	58.86	58.33	58.59	78.98	92.25	**93.01**	80.33	87.97	86.63	81.67	83.70	80.24
CMP	69.67	78.00	85.19	71.94	76.44	70.24	74.21	74.88	55.30	**89.78**	**93.77**	90.42	79.26	87.72	79.58	68.74	81.67	68.74
Mahalanobis	70.75	83.12	83.92	72.68	79.94	70.49	74.61	76.77	57.07	77.86	90.55	74.64	78.08	85.84	74.84	78.3	81.13	75.05
SanText	74.26	72.79	72.10	72.23	74.94	73.67	70.21	77.09	75.23	76.93	89.41	86.36	77.98	78.78	72.24	79.04	68.15	58.12
Gumbel	**89.64**	**88.60**	88.86	**78.26**	77.65	77.69	66.89	66.71	66.51	76.68	75.74	85.64	77.94	75.33	70.47	79.20	74.91	55.29
Vickrey	71.77	86.88	**90.75**	73.46	**83.20**	**84.27**	75.14	**79.52**	77.78	76.05	83.61	83.23	77.43	70.05	69.88	78.8	56.48	56.53
TEM	72.37	84.23	79.00	76.98	72.08	59.3	**81.59**	59.93	39.60	83.53	83.59	81.11	69.68	66.25	60.81	55.82	48.92	40.52
Dimension: 100																		
SynTF	81.19	81.99	81.94	69.46	70.28	70.36	57.73	58.56	77.00	80.24	91.07	**93.92**	**79.70**	86.99	88.40	79.16	82.90	**82.88**
CMP	65.74	78.34	87.37	69.32	77.55	75.94	72.91	76.77	64.51	**90.33**	90.29	89.33	79.68	86.52	79.19	69.03	82.75	69.04
Mahalanobis	65.99	74.44	85.64	69.58	75.01	75.75	73.17	75.59	65.85	76.18	88.40	88.11	78.44	85.67	77.09	**80.70**	82.95	66.08
SanText	80.87	79.34	77.12	**80.95**	79.54	77.19	81.02	**79.74**	77.27	75.78	89.91	87.95	77.25	79.53	76.11	78.71	69.16	64.28
Gumbel	**87.88**	**88.57**	90.51	77.49	78.03	79.52	67.09	67.48	68.53	73.22	80.7	78.47	75.65	78.94	75.26	78.07	77.19	72.04
Vickrey	66.68	84.45	88.29	70.08	**81.87**	**83.43**	73.48	79.28	**78.58**	72.16	84.03	83.36	74.81	70.32	69.95	77.45	56.61	56.55
TEM	73.35	83.26	79.07	77.60	69.41	59.48	**81.85**	55.57	39.88	85.4	81.43	80.79	70.90	63.09	60.67	56.4	44.74	40.56
Dimension: 300																		
SynTF	78.86	79.14	79.44	67.86	68.38	68.70	56.86	57.62	57.95	84.77	82.21	92.65	**82.69**	**81.72**	**88.07**	**80.60**	**81.23**	**83.50**
CMP	64.22	69.59	82.82	68.28	71.99	76.92	72.34	74.38	71.02	**88.56**	81.89	**93.83**	78.11	81.25	85.91	67.65	80.62	78.00
Mahalanobis	65.33	67.24	77.48	69.14	70.53	74.63	72.95	73.83	71.79	74.14	80.90	89.91	76.94	80.73	81.98	79.74	80.57	74.06
SanText	83.45	80.89	80.62	**82.66**	80.22	**79.17**	**81.87**	**79.55**	**77.71**	67.23	84.25	89.77	71.51	80.60	79.05	56.25	68.28	73.81
Gumbel	**87.14**	**86.56**	87.05	77.70	**77.21**	77.98	68.26	67.85	68.92	84.95	**89.84**	83.22	70.6	79.06	78.52	56.25	68.28	73.81
Vickrey	66.80	71.64	80.08	70.11	73.49	78.26	73.43	75.33	76.44	65.63	84.08	84.08	70.39	70.16	70.41	75.15	56.24	56.73
TEM	66.16	79.18	78.91	72.78	61.47	59.39	79.39	43.77	39.88	61.22	79.88	77.85	67.51	60.08	58.7	73.81	40.29	39.55

Table 8: Privacy-Utility Composite (PUC) scores.

C. Perturbation Examples

Here, we display two representative samples of perturbed text under differing parameters and mechanisms, both for an IMDb (Table 9) and AG News (Table 10) sentence.

Algorithm	d	ε	Sentence
Original:			Sorry, gave it a 1, which is the rating I give to movies on which I walk out or fall asleep. In this case I fell asleep 10 minutes from the end, really, really bored and not caring at all abour what happened next
Preproc:			sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
CMP	50	1	miniseries sphincter malaysia victory ambassador worldfest deuce hah humid parliament uranium sergei inextricably cleric genus belleau bigalow deck wistfully expectable
CMP	50	5	punish relay brunt ready channel sky fall valve face vantage recordist flicked stretch e feel iago serious family happened ford
CMP	50	10	luckily gave rating extra movie halfway fall asleep case unchanged awake minute closing really good bored unable caring happened opening
SynTF	50	1	dingy chip_in military_rank cave_in motion_picture pass devolve gone encase return gone minute_of_arc ending really genuinely drill non give_care befall adjacent
SynTF	50	5	deplorable have grade cave_in flick walk_of_life descend at_peace vitrine return asleep minute terminate real genuinely drill non handle chance adjacent
SynTF	50	10	no-count sacrifice valuation grant flick walking come_down departed pillowcase felled_seam departed second cease truly rattling drill not handle occur following
TEM	50	1	opener anxiety debilitating ingenue diminutiveaggressive lewisburns forjust macha said expressionbr lebanese chipped connors crazed marriedthis atrocitybut personification staying yuunagi geology
TEM	50	5	sorry gave rating give movie walk fall asleep case fell asleep minute end really really answer bored determined caring happened next
TEM	50	10	sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
Mahalanobis	50	1	meh fatty expectable stalag biggest hideout iscariot falkland consecutive brier baseman loch mumbai seeded ahmed mythology verifiably socorro eking emission
Mahalanobis	50	5	katrina directorial interception copy movie dedicated fashion knife related lower audible minute attendance feeling love wince always teach heimlich avenue
Mahalanobis	50	10	sorry protector rating give movie fiance craze asleep charged rose asleep minute forced getting really depressed necessarily caring happened next
SanText	50	1	predictably gave rating give movie deemed fall asleep grass fell asleep minute end devastatingly fury relating not caring happened znaimer
SanText	50	5	sorry gave expands give movie walk fall asleep case tumbled panicking scoring end so really bored not embracing feverishly next
SanText	50	10	sorry referring rating give movie approaching coincide asleep case fell woke minute end really so dreamer not nurturing happened next
Gumbel	50	1	awful picked rated need comedy walking coming waited prosecution soared woke half end always something boring could trusting happened set
Gumbel	50	5	pity took rating giving movie walking year slept case dipped suddenly superb coming sure really boring cannot caring exactly starting
Gumbel	50	10	sorry gave notch needed drama walk rise waited complaint dropped waking goal end thing lot scared nothing caring knew next
Vickrey	50	1	brokedown fwd satellite matondkar gardening spokesman blubber ej eruption readin hearkens uschi connecting schamus overly dentistry meh andalou deathly witching
Vickrey	50	5	heartache proposition score harry undertaking nap period ingrate rafter nice breech zenon sustained pretty sure suicidal earth grateful sinking first
Vickrey	50	10	stupid sunday downgrade without show cruising trend momentarily case tumbled awakens equalizer turning think think harder say physically happen scheduled
CMP	100	1	undocumented puree bernarda labor blanka buh towel gf hah roof million larceny cholera kool duster shawn homeless ankle helfgott jazz
CMP	100	5	simba attract haunted tech gaby theresa pork scary charity testimony down save nile exerting percent laced splicing simmering offer syria
CMP	100	10	sorry appeared rating trying movie fancy distort climb case northern asleep minute end probably pretend furious publicly caring surprising soon
SynTF	100	1	good-for-naught yield rank impart motion-picture_show manner_of_walking crepuscule at_peace font flow asleep minute_of_arc finish rattling real bore non wish go_on next
SynTF	100	5	gloomy throw military_rating consecrate picture walk_of_life declension numb causa strike_down numb moment end rattling very drill non worry materialize succeeding
SynTF	100	10	sad hold blackleg cave_in flick base_on_balls decline asleep caseful settle benumbed second cease genuinely genuinely bore non deal pass_off following
TEM	100	1	vandal kikkis bijomaru moliere betteralso crandall inmatesbr rightbr humped xvichiasergo midsummer keyboardist ferry mainlybecause crony sarlac overriding grossout drummed flopbecause
TEM	100	5	sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
TEM	100	10	sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
Mahalanobis	100	1	interpol disposable stuffing withstand cyclonic bardot dah ira hah unforced oom budget pulp saucepan simmer data srebrenica dna semitism diced
Mahalanobis	100	5	harry beheading restrictive und loos mal demolition subsistence proxy mountain blocking sung violently receiver concept diol buckingham leporidae written language
Mahalanobis	100	10	forgot opening rating give movie walk asleep attorney fell asleep minute ceiling code really deliriously story harrowing happened search
SanText	100	1	sorry ferro rating give truffaut walk fall asleep case fell asleep minute end really deutschland bored not afro paved next
SanText	100	5	sorry unconquerable rating conducting movie walk fall asleep concludes fell asleep minute end really musing bored not caring surly next
SanText	100	10	thankful this rating give upcoming lined fall asleep case slipped woke minute end really really bored not caring why next
Gumbel	100	1	thank gave disapproval take blockbuster sit rise woke trial slid woke minute start maybe really jaded so nurturing happened week
Gumbel	100	5	glad gave rating give hollywood walk end drunk case slipped asleep equalizer time everyone thought bored not sick happened day
Gumbel	100	10	ok brought disapproval come drama walked coming crawled complaint tumbled crawled kick start something think bored would caring knew start
Vickrey	100	1	constrict stir firefight jc elimination exhibitor ali birdman kissed imtiaz clothing gymnastics mile inexhaustible traveler meh decency mil album neagle
Vickrey	100	5	happy possession rating right tinge ball colombian toddler suburb fell gawk heartedly host guy fascinating bad escape truer ya unrealized
Vickrey	100	10	frankly later nielsen grab soundtrack walking since squirming relation gained wandered rush gap rely something downright believe widowed everybody time
CMP	300	1	gr gft expectable chakra grandparent gored magritte noo sniper breakfast meh substantive paternal verifiably viking flute erm striker muddies shane
CMP	300	5	relay neighborhood crime dubai hiroshima vampyre sandal kilcher estimate evers studiously mib cowgirl puzzle coldest stv danube berkley pulitzer del
CMP	300	10	cardiff interpretation efficiency hollywood movie stooped reacting sleep case export asleep minute dillon unable lot bored interfere trainor depressive gunter
SynTF	300	1	meritless devote rat ease_up flick walkway fall departed character diminish asleep mo cease actually in_truth blase not deal materialize future
SynTF	300	5	disconsolate afford rat devote movie base_on_balls downfall gone vitrine fell departed narrow terminal in_truth truly tire non lovingness take_place future
SynTF	300	10	bad cave_in blackleg ease_up movie paseo hang at_peace font precipitate asleep mo stop truly actually bore not manage happen succeeding
TEM	300	1	peckenpahs urchin iñárritu lansburys streaming clout goosebump kissed welcomed maggies iwhippedi waswell occuped damme unbielevable calligraphy cameraman nula sharie british
TEM	300	5	sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
TEM	300	10	sorry gave rating give movie walk fall asleep case fell asleep minute end really really bored not caring happened next
Mahalanobis	300	1	perspicacious peace gf pellet gomer gargan raspberry kursk no prime wisconsin pickier reddin salvific designer clunkers incursion martyr hurd umm
Mahalanobis	300	5	sincere ha hrithik goalkeeper batman innes pole astral bellucci visa disfigured clan wale geometry faceoff simon sharia humperdink von faulty
Mahalanobis	300	10	sad summary rating age movie mum monarch asleep psychiatric fell asleep goalkeeper talker silver prototype improvising office caring thigpen declined
SanText	300	1	sorry gave rating give granddaughter walk fall asleep case fell asleep minute end really workhorse generate agreeing caring happened next
SanText	300	5	sorry gave rating give movie walk fall asleep hypothetically huggable selina memorably end crucially really starstruck insisting caring happened next
SanText	300	10	sorry gave rating give movie begun fall asleep concluded fell asleep minute end really bored be caring happened next
Gumbel	300	1	embarrassed give indicating gave filmed anyway though awake reason dropped woke minute end definitely certainly tired though evera happen week
Gumbel	300	5	disappointed put rating giving movie walking fall awake example falling asleep equalizer ended really know bored although elderly happened expected
Gumbel	300	10	ashamed giving rating given starred walked coming sleep example slid fortunately came however obviously certainly bored be loving happening take
Vickrey	300	1	dah mayoral herein wachowski address ee corneau blazing ketchup observatory curled verdi thematic zen materialises ishwar wrestlemania nicholsons sonja interference
Vickrey	300	5	pepe pota eavesdrops hatching stunt yeop traumatizing takoma detained factly picher hitch light englund encyclopedia glanced calcium ditzy pasta chromosome
Vickrey	300	10	miserable assertion plunging invocation jerker sabre competing appetizer homicide dated suspended sanchez levy go consistency scene entertained flawed dreamt cbs

Table 9: Example text output on the IMDb dataset.

Original:			Olympic Baseball Makes a Hit in Greece ATHENS (Reuters) - The team that did near the worst in Olympic baseball drew the most fans while the team among the most feared were not there
Preproc:			olympic baseball make hit greece athens reuters team near worst olympic baseball drew fan team among feared

Algorithm	d	ε	Sentence
CMP	50	1	kazakhstani billboard physically humor debian pour mcconnaughey steller excavation orthodox moosehead fax forehand maggette republican hristo hornet
CMP	50	5	olympics mlb expensive smash jerez quote reuters valuable nablus wreaked olympics legion character era buyout mountainous believe
CMP	50	10	olympic baseball whatever hit greece athens reuters team northern worst olympic baseball stinging fan team among ridden
SynTF	50	1	Olympian baseball urinate pip Hellenic_Republic Athinai reuters team_up virtually worst Olympian baseball_game pull devotee squad among fear
SynTF	50	5	Olympic baseball shuffling pip Ellas Greek_capital reuters team close defective Olympic baseball thread buff squad among reverence
SynTF	50	10	Olympic baseball clear shoot Greece capital_of_Greece reuters squad virtually high-risk Olympic baseball thread devotee team among revere
TEM	50	1	cornerback mayer ampts trekker pariah rsen macrozonaris menu separately mae barren liven soreness assigning bayonne dts seat
TEM	50	5	olympic baseball make hit greece athens reuters team near worst olympic baseball drew fan team among feared
TEM	50	10	olympic baseball make hit greece athens reuters team near worst olympic baseball drew fan team among feared
Mahalanobis	50	1	st philippe rangana touchdown metway corrects mutate bernabeu deivi violence megabit liossia damp pheidippides maumoon disorder prehistoric
Mahalanobis	50	5	ioa hockey britain strip republic powerpoint mtfg evolution born hit oly specific inspirational equivalent human numerous outplays
Mahalanobis	50	10	olympic baseball make hit greece athens reuters team near worst olympic baseball drew drew team among causing
SanText	50	1	olympic baseball make hit follows athens reuters disposed near eserver amir baseball underscore fan team among untangle
SanText	50	5	olympic baseball make hit greece athens reuters team where dismal skating baseball drew fan team among feared
SanText	50	10	olympic baseball make hit greece athens reuters team near aftermath skating coaching drawing fan team among feared
Gumbel	50	1	olympic mlb could drove cyprus canberra posted team outside faced competition nba drawing buzz team well threatened
Gumbel	50	5	olympic yankee need another cyprus melbourne correspondent team northwest wake world nba responded joke squad represent blamed
Gumbel	50	10	champion sox making hit greece athens reporting team outside hardest olympic sox drawn sensation league especially feared
Vickrey	50	1	upload honorable mwai retake sternest xu amp trillion nazi bruin hayat publisher hitler quarterback mazda westbound amazon
Vickrey	50	5	berth baseball sheer storm hiptop gymnastic analysis shield hospital complete musharraf financed slain frankfurt playoff hartford professional
Vickrey	50	10	olympics slugger easier hitting bulgaria athens quoting coaching southwest devastating olympics hockey mccarthy mighty squad well politician
CMP	100	1	ligament al vote domecq balco mathematical amf minority misdemeanor anbar damarcus gnome veterinary seasoning dilshan dugout khost
CMP	100	5	bafta yawkey told unflattering javier panathinaikos poll team burst cheater rogge colorado gamespot fan corrects openly eritrea
CMP	100	10	olympic baseball come hit greece athens reuters team near mexico decider baseball break contributing team among feared
SynTF	100	1	Olympic baseball_game induce hit Ellas Athens reuters team come_on high-risk Olympic baseball tie winnow team_up among revere
SynTF	100	5	Olympian baseball_game establish striking Hellenic_Republic Athens reuters team_up skinny sorry Olympic baseball_game withdraw lover team among reverence
SynTF	100	10	Olympic baseball_game wee striking Greece Athens reuters squad virtually regretful Olympian baseball draw_and_quarter winnow team among fear
TEM	100	1	relaxation winnipegthree bermittelt tremor ballet recouped diameter hualien belgrade flutie mom visit workplace flexitronics ad violently
TEM	100	5	olympic baseball make hit greece athens reuters team near worst olympic baseball drew fan team among feared
TEM	100	10	olympic baseball make hit greece athens reuters team near worst olympic baseball drew fan team among feared
Mahalanobis	100	1	estuary condor uh indignity average reactor heatwave de smoking mideast threw unbeaten chesnot assateague alitalia lethal vaccine
Mahalanobis	100	5	origin softball verbal andhra collider bayern bad raced container binge olympic haired orleans biker suburb including dam
Mahalanobis	100	10	medalist mlb reach hit greece athens reuters team near cisco olympic baseball outing happier team often hearing
SanText	100	1	olympic selfish make hit shatter athens flagged team near worst bennger murkowski drew fan team among feared
SanText	100	5	olympic baseball it willis greece athens reuters team near worst olympic walking pushing lightning latter among feared
SanText	100	10	olympic baseball make throw greece athens thomson coaching nearby worst olympic baseball drew fan team among feared
Gumbel	100	1	swimming softball put coming bulgaria athens reporting team near aftermath medalist league came fan team leading believed
Gumbel	100	5	olympics softball need went albania olympic wsj team near worst olympic mlb attention spectator player especially feared
Gumbel	100	10	olympics baseball making went iceland nagano reuters team along dismal olympics baseball turned buzz soccer many feared
Vickrey	100	1	seung arabiya malik najaf pope sistani crore douri taiex retina stimulant dictator hristo muqtada awami du accurately
Vickrey	100	5	medal union siebel drug stealth chapter giant roadside repairing sluggish olympic optioned uruguay garden amateur elderly jolted
Vickrey	100	10	marathon streak kind drove italy sydney bloomberg retired town disruption relay nhl matt villain squad involving connection
CMP	300	1	paavo antivirus petronas peirsol yielding sodomy topix axa scandal java animated narcotic clearinghouse transmeta sega bush bnk
CMP	300	5	herbold rockies rutherford contactless greece pension healthday nfl lunar easyjet micronesia midriff marvel fan nvidia disarming altoona
CMP	300	10	aid hitter make hit greece intl cote philadelphia shattering worst surpassed baseball merz brawl car muslim shunned
SynTF	300	1	Olympic baseball lay_down gain Greece capital_of_Greece reuters team nearly unsound Olympian baseball sop_up fan team_up among fear
SynTF	300	5	Olympic baseball score smasher Hellenic_Republic Athens reuters squad close uncollectible Olympian baseball John_Drew lover team_up among venerate
SynTF	300	10	Olympic baseball give bang Hellenic_Republic capital_of_Greece reuters squad virtually sorry Olympic baseball_game Drew rooter team_up among reverence
TEM	300	1	furless resuming edgy eyepods kazmirs senseis approving ederal physical benefited impose oreilly kastors hobbyist coxnyse found berkleigh
TEM	300	5	olympic baseball make hit greece athens reuters team near worst olympic baseball drew fan team among feared
TEM	300	10	olympic baseball make hit greece athens reuters team near worst olympic baseball drew fan team among feared
Mahalanobis	300	1	heraldic torrential fingerprint severe tamoxifen newspoll racketeer varejao jemaah marwan mental platinum atlanta centerfold incurable grudzielanek odumbe
Mahalanobis	300	5	accreditation shutout scientific ibf cide ekimov anxiously alberta unhurt overcapacity millionth fab espn marrero pleaded trampled terminal
Mahalanobis	300	10	olympic favorite move calculate alertness victor google team near fracture fired obscenity catalanotto snapping bubble violence curt
SanText	300	1	olympic baseball make hit greece wheelchair reuters team near description olympic baseball drew fan concerned among toting
SanText	300	5	placing baseball make touched greece athens packer offseason near worst advocating mlb drew fan team among feared
SanText	300	10	nagano mlb willing hit greece alerted reuters fourteen near devastating olympic baseball drew fan coached among feared
Gumbel	300	1	nagano baseball making hit greece olympics reuters time near terrible olympic baseball drew natsha team others fearing
Gumbel	300	5	olympics football even hit bulgaria athens rgentina squad close seen olympics baseball drew fan football although feared
Gumbel	300	10	olympic hockey even followed greece greece thomson squad nearby worst event mlb drew fan finished also possibly
Vickrey	300	1	babson skywatchers bared topix belmarsh mont abg yngling beatify nonfarm shoaib square felipe hindu production expression truce
Vickrey	300	5	streamcast linkedin ggp naples bomb mcauliffe implant michigan prairie poorer daisuke amr empire apiece que garland seasonal
Vickrey	300	10	decathlon devote lend critically piraeus janeiro cameraman castro southeast worst nagano hockey drawn bookstore season surging distracted

Table 10: Example text output on the AG News dataset.