Supplementary Information

Comparative Document Summarisation via Classification

This supplementary information contains additional information required for the main paper, including details of greedy and gradient optimisation, analysis of crowdsourced evaluations and results table of automatic evaluations. The source code and datasets of this work is available online. [1]

Contents

A Greedy Algorithm

The greedy approach to maximise a utility function $\mathcal{U}(\bar{\mathbf{X}})$ is outlined in Algorithm 1. We iterate through all groups, adding one point to the summary at a time, selected as one that brings the largest marginal gain to the utility function $\mathcal{U}(\cdot)$. Details necessary for computing the marginal gain in step 7 are found in § B.

Algorithm 1 Greedy algorithm to maximize objective $\mathcal{U}(\cdot)$

Require: $\{\mathbf{X}_{1:G}\}$: Groups of documents,

1: G: number of groups, M: number of prototypes per group
2: **procedure** GREEDYMAX($\mathbf{X}, G, M$)
3: $(\forall g \in 1 \dots G)\, \bar{\mathbf{X}}_g \leftarrow \{\}$
4: **for** m from 1 to M **do**
5: $\bar{\mathbf{X}} \leftarrow \cup_{g=1}^{G} \bar{\mathbf{X}}_g$
6: **for** g from 1 to G **do**
7: $\bar{\mathbf{x}}_{g,m} \leftarrow \underset{\mathbf{x}_g \in \mathbf{X}_g \setminus \bar{\mathbf{X}}_g}{\operatorname{argmax}} \Delta_{\mathcal{U}}(\mathbf{x}_g | \bar{\mathbf{X}})$
8: $\bar{\mathbf{X}}_g \leftarrow \bar{\mathbf{X}}_g \cup \bar{\mathbf{x}}_{g,m}$
9: $\bar{\mathbf{X}} \leftarrow \cup_{g=1}^{G} \mathbf{X}_g$
10: **return** $\bar{X}$

[1] https://github.com/computationalmedia/compsumm

B Gradients and Discrete Derivatives of the MMD objectives

The equation and gradient gradient of $MMD^2(\bar{\mathbf{A}}_g, \mathbf{X}_g)$ for the RBF Kernel are defined as:

$$MMD^2(\bar{\mathbf{A}}_g, \mathbf{X}_g) = -\frac{2}{M \times N_g} \sum_{i=1}^{N_g} \sum_{j=1}^{M_g} k(\bar{\boldsymbol{a}}_{g,j}, \mathbf{x}_{g,i}) + \frac{1}{M_g^2} \sum_{i,j=1}^{M_g} k(\bar{\boldsymbol{a}}_{g,i}, \bar{\boldsymbol{a}}_{g,j}) \tag{1}$$

$$\forall_{l \in 1 \ldots M_g} \nabla_{\bar{\boldsymbol{a}}_{g,l}} MMD^2(\bar{\mathbf{A}}_g, \mathbf{X}_g) = \frac{4\gamma}{M_g} \Bigg(-\frac{1}{N_g} \sum_{i=1}^{N_g} k(\bar{\boldsymbol{a}}_{g,l}, \mathbf{x}_i)(\mathbf{x}_i - \bar{\boldsymbol{a}}_{g,l}) + \frac{1}{M_g^2} \sum_{i=1}^{M_g} k(\bar{\boldsymbol{a}}_{g,i}, \bar{\boldsymbol{a}}_{g,l})(\bar{\boldsymbol{a}}_{g,i} - \bar{\boldsymbol{a}}_{g,l}) \Bigg) \tag{2}$$

$MMD^2(\mathbf{A}_g, \mathbf{X}_{\neg g})$ can also be computed in a similar way, by replacing $\mathbf{x}_{g,i}$ by $\mathbf{x}_{\neg g,i}$ in equation (2), this will yield the objective of $\mathcal{U}_{\textit{diff}}(\bar{\mathbf{X}})$ (5). The first term of the equation (2) corresponds to the gradient of first term of equation (1). Hence, it will yield the gradient of the objective $\mathcal{U}_{\textit{div}}(\bar{\mathbf{X}})$ (6).

Let V_g be the indices of $\mathbf{X}_g$ and S_g be the indices of $\bar{\mathbf{X}}_g$. The discrete derivatives for $-MMD^2(\bar{X}_g, X_g)$ is.

$$\begin{aligned}\Delta_{-MMD^2(\bar{\mathbf{X}}_g, \mathbf{X}_g)}(\mathbf{x}_g \mid \bar{\mathbf{X}}_g) = \frac{1}{|S_g|+1} \Bigg(& \frac{2}{|V_g|} \sum_{i \in V_g} k(\mathbf{x}_i, \mathbf{x}_g) - \frac{2}{|V_g||S_g|} \sum_{i \in V_g, j \in S_g} k(\mathbf{x}_i, \mathbf{x}_j) + \\ & \frac{2|S_g|+1}{|S_g|^2(|S_g|+1)} \sum_{i,j \in S_g} k(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{(|S_g|+1)} \sum_{i \in S_g} k(\mathbf{x}_i, \mathbf{x}_g) - \frac{1}{|S_g|+1} k(\mathbf{x}_g, \mathbf{x}_g) \Bigg)\end{aligned} \tag{3}$$

Discrete derivatives of different MMD objectives (equations 5, 6) can be built upon equation (3). The discrete derivative of λ term in equation (6) is given by first two terms of equation (3). The Discrete derivative of equation (2) can be computed in a similar way. Equations for discrete derivatives allow greedy optimisation (§A) to be done efficiently.

C Methodology for data collection

Topic curation. We curate an initial list of 10 topics in June 2017 that satisfying vthe criteria of having non-trivial news coverage and being controversial. In this work, we use *Beef Ban*, *Capital Punishment* and *Gun Control* topics. The other topics are *Climate change, Illegal immigration, Refugees, Gay marriage, Animal testing, Cyclists on road* and *Marijuana*. We want to focus on controversial topics since they are likely to be discussed in the future since their coverage lasts for a long time. Controversy is an important topic for research in social media and online political discourse, is also important in real-world applications such as intelligence and business strategy development.

To obtain various opinions on contemporary social problems, we choose Twitter as a source since it is frequently used for reporting and sharing related news articles. Garimella et al. (2018) use similar approach generating Twitter dataset on the controversial topics. The authors consider Twitter hashtags as query and use similarity function to retrieve similar hashtags. We obtain embedded news articles from Twitter posts to generate a dataset and use a different expansion approach to retrieve related hashtags.

Topic	Queries	#Tweets	#News	#News (cleaned)
Beef Ban	beef ban, beefban	304,234	17,131	1,543
Capital Punishment	death penalty, deathpenalty, capital punishment	11,052,295	66,542	7,905
Gun Control	gun control, guncontrol, gunsense, gunsafety, gun laws, gun violence	36,533,525	130,312	6,494

Table 1: Controversial Topic Dataset Statistics

Query curation. We use a hashtags expansion approach (Verkamp and Gupta, 2013) to curate relevant queries for each topic. We first manually select a single query for each topic, then use it to collect Twitter posts for two weeks. These posts are used as an initial data set that we create a query set based on. We extract the 10 most common hashtags that appear in the initial dataset. These hashtags are used to query the same dataset again and then we re-extract the 10 most common hashtags from the query result. We continue this iteration several times until the hashtags used for query and the re-extracted hashtags are the same. All of the topics finish generating a query set after 4~5 iterations.

Method	#unique workers	correct by majority	correct judgements
kmeans	31	81	240
mmd-diff-grad	25	94	270
nn-comp-greedy	28	80	243
mmd-diff-greedy	29	83	235
Total	40	126	378

Table 2: Results of Human Pilot Study on Classification Task. Unique workers participating in classifying test articles for each method is in first column. Correct by majority means the number of test articles (out of 126) classified correctly by majority (at least two people). Correct Judgments indicates the number of individual judgments that are correct (out of 378)

Based on the query set generated using the hashtags expansion, we perform additional filtering. Location hashtags such as #Florida or #Alabama are removed to prevent detailed locations being discussed. Some hashtags like #cow, #beef, #PJNET, and #2A are excluded since they are not directly related to the topics or are too general. As a result, *Beef Ban* topic is defined by a single query while *Capital Punishment* and *Gun Control* include more diverse hashtags in the query set. Table 1 summaries the query set used for each topic.

Article extraction. After generating a query set for each topic, we fetch the Twitter stream that includes any of the hashtags in the query set. Twitter post frequently includes embedded news articles related to the post. We focus on the news articles in this work since they generally include more coherent stories than the corresponding Twitter post dataset. We extract the embedded news articles by visiting the article URL and downloading the content from it. By doing that, we can collect news articles that are mentioned and shared in ongoing social media which can be a measure of how important and accurate the news is. We clean the data by filtering spam articles and removing duplicate articles mentioned in multiple tweets. To increase the relevance, we remove garbage texts such as "Subscribe to our channel", "Please sign up" or "All Rights Reserved" that repeatedly appear with the news content. Table 1 reports the number of the Tweets and the news articles before and after the cleaning.

D Human Evaluation Results

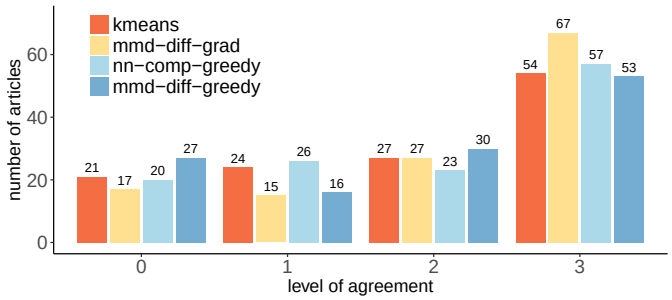

Figure 1: Shows the number of test articles that humans correctly classified, at different agreement levels. A level of 3 means all participants correctly classified the test article, while 0 means all participants incorrectly classified the test article.

Table 2 shows the number of unique participants and the number of correct judgements for each methods from the human evaluation result. The number of unique participants answering test questions ranged from 25 to 31. The union of the participants involved in any four methods is 40. The number of articles correctly classified by the participants when we evaluate by majority voting and treat each judgment shows the efficacy of the proposed method *mmd-diff-grad* over other methods including *kmeans* baseline.

Figure 1 shows the level of agreement across participants for each method. First we note that participants were frequently able to complete the task of classifying new articles correctly into one of two groups, this is shown by the large fraction of articles for which the correct group was unanimously chosen. Compared to other comparative prototype selection methods *mmd-diff-grad* has the largest number of articles correctly classified by all three participants, beating the next best *nn-comp-greedy* by 10 articles. Consequently *mmd-diff-grad* also has fewer articles which were unanimously assigned to the incorrect group by participants.

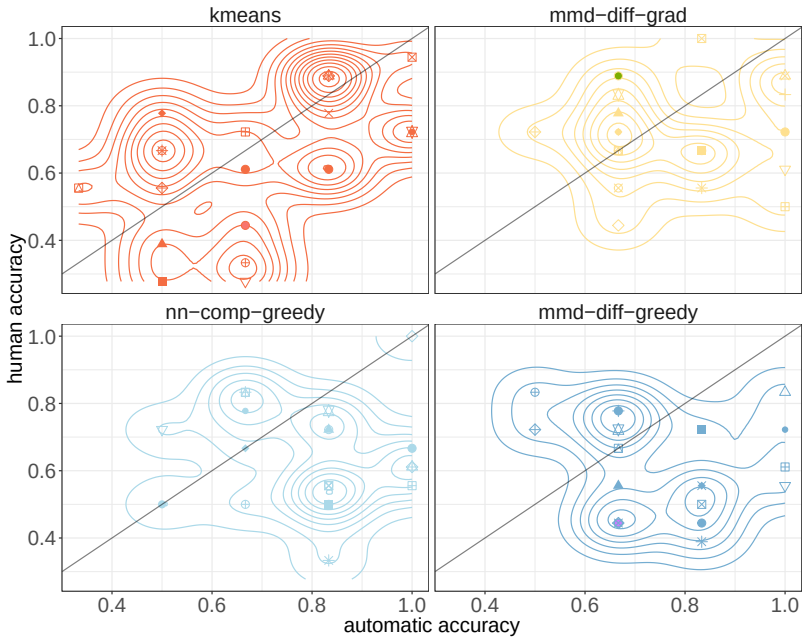

Figure 2: Density plot of human accuracy vs. machine classifier (SVM) accuracy. Each point means a pair of summaries and we average the accuracy over the 6 test articles. People are doing roughly same as SVM for *kmeans* and *mmd-diff-grad*, whereas people seems to do worse than SVM in case of greedy methods.

Figure 2 shows the density plot comparing human accuracy vs. SVM accuracy in classifying small set of test articles we had for human evaluation.

E Results Table

Automatic evaluation results table corresponding to error-bar plot in Figure 4 is given in Tables 3, 4, 5 and 6, We report the mean of balanced accuracy and 95% confidence interval of the 10 random splits. The left four columns use 1-NN and the right four use SVM. We report results on 2, 4, 8, or 16 prototypes per group. The highest performing method is in bold and second highest performing one is in italics.

method	2	4	8	16	2	4	8	16
kmedoids	0.501 ± .010	0.505 ± .011	0.528 ± .012	0.530 ± .021	0.498 ± .004	0.501 ± .007	0.498 ± .004	0.494 ± .014
kmeans	0.510 ± .017	0.542 ± .012	*0.546 ± .007*	0.576 ± .008	0.513 ± .016	*0.539 ± .010*	0.541 ± .007	0.543 ± .007
mmd-critic	0.499 ± .003	0.514 ± .016	0.515 ± .013	0.531 ± .009	0.498 ± .005	0.514 ± .014	0.512 ± .022	0.531 ± .014
mmd-diff-grad	*0.534 ± .011*	*0.538 ± .014*	0.540 ± .020	**0.582 ± .010**	0.525 ± .018	0.538 ± .015	**0.559 ± .008**	0.568 ± .011
mmd-div-grad	**0.539 ± .013**	**0.545 ± .008**	**0.564 ± .014**	*0.579 ± .011*	0.523 ± .013	0.548 ± .015	0.556 ± .011	0.566 ± .008
nn-comp-greedy	0.509 ± .011	0.515 ± .008	0.544 ± .009	0.577 ± .007	0.512 ± .017	0.524 ± .015	0.556 ± .011	**0.572 ± .010**
mmd-diff-greedy	0.530 ± .009	0.536 ± .012	0.545 ± .013	0.564 ± .013	*0.533 ± .012*	**0.555 ± .012**	*0.557 ± .009*	0.567 ± .007
mmd-div-greedy	0.530 ± .010	0.525 ± .011	0.539 ± .012	0.563 ± .009	**0.538 ± .009**	0.535 ± .009	0.547 ± .009	*0.571 ± .010*

Table 3: Classification performance on *Capital Punishment* News dataset. (left) 1-NN, (right) SVM.

method	2	4	8	16	2	4	8	16
kmedoids	0.592 ± .016	0.586 ± .011	0.582 ± .025	0.589 ± .023	0.524 ± .031	0.514 ± .019	0.542 ± .018	0.555 ± .035
kmeans	0.582 ± .020	**0.613 ± .017**	*0.622 ± .030*	0.629 ± .025	0.564 ± .022	0.592 ± .019	0.577 ± .029	0.579 ± .021
mmd-critic	0.541 ± .031	0.524 ± .028	0.530 ± .025	0.535 ± .033	0.543 ± .030	0.537 ± .023	0.513 ± .038	0.528 ± .027
mmd-diff-grad	*0.595 ± .019*	0.587 ± .023	0.610 ± .017	*0.633 ± .020*	**0.594 ± .019**	0.603 ± .018	0.617 ± .026	0.640 ± .031
mmd-div-grad	**0.602 ± .021**	*0.605 ± .015*	0.605 ± .028	**0.636 ± .028**	*0.591 ± .020*	*0.604 ± .021*	0.614 ± .023	**0.648 ± .031**
nn-comp-greedy	0.587 ± .018	0.600 ± .027	**0.624 ± .029**	0.627 ± .026	*0.591 ± .027*	**0.615 ± .019**	**0.628 ± .017**	0.640 ± .023
mmd-diff-greedy	0.592 ± .027	0.595 ± .021	0.615 ± .019	0.629 ± .021	*0.591 ± .028*	0.594 ± .033	*0.619 ± .023*	0.638 ± .020
mmd-div-greedy	0.586 ± .014	0.578 ± .021	0.593 ± .022	0.616 ± .025	0.579 ± .029	0.581 ± .027	0.600 ± .021	*0.644 ± .020*

Table 4: Classification performance on *Beefban* News dataset. (left) 1-NN, (right) SVM.

method	2	4	8	16	2	4	8	16
kmedoids	0.501 ± .016	0.504 ± .012	0.518 ± .016	0.518 ± .013	0.500 ± .000	0.500 ± .001	0.498 ± .008	0.506 ± .007
kmeans	0.505 ± .009	0.506 ± .006	**0.538 ± .013**	0.542 ± .014	0.506 ± .009	0.504 ± .009	0.509 ± .011	0.510 ± .009
mmd-critic	0.506 ± .006	0.511 ± .013	0.507 ± .011	0.514 ± .014	0.505 ± .007	0.503 ± .011	0.511 ± .013	0.521 ± .014
mmd-diff-grad	**0.531 ± .009**	**0.529 ± .006**	0.532 ± .008	**0.566 ± .006**	**0.538 ± .008**	*0.534 ± .015*	**0.541 ± .010**	0.546 ± .012
mmd-div-grad	0.525 ± .011	*0.525 ± .009*	*0.537 ± .010*	*0.563 ± .011*	*0.535 ± .014*	**0.538 ± .010**	0.538 ± .013	*0.549 ± .011*
nn-comp-greedy	0.502 ± .010	0.518 ± .011	0.535 ± .014	0.555 ± .009	0.515 ± .014	0.523 ± .012	0.521 ± .007	0.537 ± .007
mmd-diff-greedy	*0.524 ± .015*	*0.520 ± .013*	0.521 ± .013	0.537 ± .010	0.512 ± .013	0.523 ± .016	*0.538 ± .007*	**0.552 ± .010**
mmd-div-greedy	0.517 ± .012	0.515 ± .010	0.519 ± .012	0.532 ± .011	0.509 ± .014	0.525 ± .012	0.532 ± .008	0.533 ± .013

Table 5: Classification performance on *Gun Control* News dataset. (left) 1-NN, (right) SVM.

method	2	4	8	16	2	4	8	16
kmedoids	0.805 ± .010	0.836 ± .014	0.862 ± .008	0.881 ± .008	0.783 ± .012	0.838 ± .019	0.864 ± .010	0.878 ± .011
kmeans	**0.823 ± .012**	**0.866 ± .010**	*0.888 ± .006*	0.909 ± .009	0.823 ± .011	*0.868 ± .010*	**0.896 ± .009**	0.911 ± .008
mmd-critic	0.560 ± .019	0.700 ± .016	0.777 ± .013	0.839 ± .010	0.362 ± .040	0.416 ± .056	0.552 ± .033	0.798 ± .023
mmd-diff-grad	*0.811 ± .011*	0.852 ± .008	0.882 ± .007	*0.910 ± .010*	**0.834 ± .012**	**0.872 ± .011**	0.889 ± .012	*0.912 ± .011*
mmd-div-grad	0.806 ± .010	0.849 ± .010	0.876 ± .007	0.907 ± .010	*0.832 ± .011*	**0.872 ± .012**	*0.892 ± .012*	**0.913 ± .012**
nn-comp-greedy	0.800 ± .011	*0.859 ± .010*	**0.890 ± .009**	**0.914 ± .010**	0.797 ± .011	0.853 ± .011	0.891 ± .009	**0.913 ± .008**
mmd-diff-greedy	0.783 ± .011	0.835 ± .009	0.871 ± .009	0.898 ± .010	0.795 ± .009	0.849 ± .011	0.880 ± .009	0.909 ± .008
mmd-div-greedy	0.784 ± .013	0.840 ± .010	0.866 ± .010	0.898 ± .007	0.798 ± .010	0.852 ± .011	0.878 ± .011	0.904 ± .009

Table 6: Classification performance on *USPS* dataset. (left) 1-NN, (right) SVM.

References

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying controversy on social media. *Transactions on Social Computing*.

Verkamp, J.-P., and Gupta, M. 2013. Five incidents, one theme: Twitter spam as a weapon to drown voices of protest. In *USENIX Workshop on Free and Open Communications on the Internet*.