Evaluation of Personalized Summarization w.r.t Temporal Variance

Student Name: Darsh Rank
Enrollment ID: **201901247**
B. Tech. Project (BTP) Report
BTP Mode: ***On Campus***
Dhirubhai Ambani Institute of ICT (DA-IICT)
Gandhinagar, India
201901247 [at] daiict.ac.in

Mentor's Name: Prof. Sourish Dasgupta
Dhirubhai Ambani Institute of ICT (DA-IICT)
Near Indroda Circle
Gandhinagar 382007, India
sourish_dasgupta [at] daiict.ac.in

***Abstract*—The objective of this research is to show that assessing the personalization of a summarization model is different from evaluating its accuracy. The accuracy-based evaluation metrics such as ROUGE do not consider the temporal variance of the user while evaluating the personalized summarization model. The study aims to introduce a new metric $e - DINS_{TV}$ which considers the rate of change of model generated summaries when compared to rate of change of user profiles with respect to time. The perception of saliency is subjective and in case of temporal variance, it changes overtime for each user.**

***Index Terms*—adaptive summarization, evaluation metric, personalization, temporal variance**

I. Introduction

Various metrics assess the quality and accuracy of a summary model [1], primarily focusing on readability. However, none of these metrics consider the model's ability to capture changes in user preference over time (temporal variance), which is crucial for measuring personalization. To evaluate how well a summarization model can generate personalized summaries that adapt to users' changing preferences, it is important to measure the degree of personalization. The summarizer's speed is also crucial in this context, as it must quickly adapt to users' shifting interests after an indefinite amount of time.

In paper [2], which was Exdos [3] based personalized summarization model were measured in terms of iterative convergence. Here convergence happens when the user stop giving feedback on the same document.

Microsoft proposed PENS framework[4], which can generate personalized headlines based on user profiles. Headlines can be considered as personalized summaries. To evaluate their model ROUGE was calculated between user-written personalized summary and model-generated personalized summary. We argue that ROUGE is an accuracy measure, it cannot measure the degree of personalization. Fig. 2 (a) shows that the model has high accuracy since the user profile and summary have a small difference, indicating that the model captures shift in user's interests quite well. However, personalization is low since their summary pair has less difference compared to the user profile pair. Fig. 2 (b) shows that the model has high accuracy while personalization is also high as the summary and the user profile pair have almost equal differences. So a model may have high accuracy but still have a low degree of personalization. Thus, accuracy measures are not adequate for measuring the degree of personalization.

The accuracy-centric measure focuses on the difference between user-written and model-generated summaries (shown by the red circle in fig 2). In contrast, the personalization-centric measure focuses on the relationship between model-generated summaries and user profile over time to evaluate the personalization capability of a summarization model.

II. Learning Outcomes

While developing an evaluation metric for personalized text summarizers, I had several learning outcomes, some of which are:

- Understanding of Natural Language Processing
- Understanding of summarization techniques
- Understanding of evaluation metrics
- Ability to evaluate and analyze results
- Development of critical thinking skills

III. Contributions

Our contributions are as follows:

- Demonstrated that evaluating a summarizer model's personalization is not the same as evaluating its accuracy.
- Proposed evaluation metric to measure the degree of personalization w.r.t temporal variance.

IV. Proposed evaluation measure

In this section, we'll go over how our proposed evaluation measure was designed.

A. Defining Insensitivity

Given document $\mathbf{D}$ and a closely similar document $\mathbf{D}^{'}$, and user profile details $\mathbf{u}$, let summarization model $M_{\boldsymbol{\theta},\mathbf{u}}$, generate the best estimated personalized summary $\widehat{S}$

$M_{\boldsymbol{\theta},\mathbf{u}} : \mathbf{D}, \mathbf{u} \mapsto \widehat{S}_u$

Given the user-profile of the same user at two different time-points, $\mathbf{u}_{(i,t)}$ and $\mathbf{u}_{(i,t+\Delta)}$ respectively, a summarization

		Manual Evaluation	Automated Evaluation	
			Reference-based	Reference-free
Accuracy		1. Factoid (2003) 2. Pyramid (2004) 3. SEE (2003)	1. Cosine similarity (1989) 2. BLUE (2002) 3. ROUGE (2004) 4. Unit overlap (2002) 5. Latent-based (2009) 6. Semi-automated pyramid (2018) 7. Automated pyramid (2017)	1. KL divergence (1959) 2. Jensen–Shannon divergence (1991) 3. FRESA (2013) 4. SummTriver (2018) 5. Summary likelihood (2013) 6. SUPERT (2020)
Quality		1. DUC 2005 readability 2. TAC 2008 readability	Sum-QE(2019)	
Degree of personalization	w.r.t temporal variance			e-D-INS$_{TVar}$ (In progress)

Fig. 1. Different evaluation metrics[1]

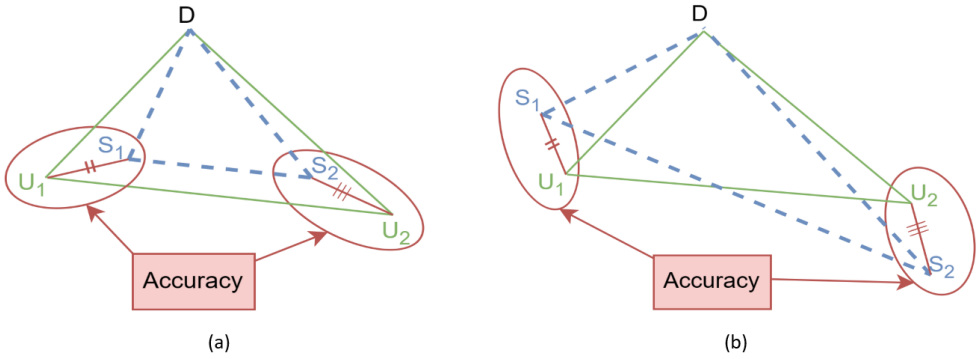

Fig. 2. (a) shows accuracy is high but personalization is low Accuracy. (b) shows both accuracy and personalization are high

model $M_{\boldsymbol{\theta},\mathbf{u}}$ is (weakly) *Insensitive-to-Temporal-Variance* (Ins-Temp-Var) iff $\forall\,(\mathbf{u}_{(i,t)}, \mathbf{u}_{(i,t+\Delta)})$, $\ni f^U_{dist}(\mathbf{u}_{(i,t)}, \mathbf{u}_{(i,t+\Delta)})^* > \tau^U_{max}$, $\mathbf{D} \approx \mathbf{D}^{'}$, and $\Delta > \tau^t_{max}$:

$$f^S_{sim}(M_{\boldsymbol{\theta},\mathbf{u}}(\mathbf{D}, \mathbf{u}_{(i,t)}), M_{\boldsymbol{\theta},\mathbf{u}}(\mathbf{D}^{'}, \mathbf{u}_{(i,t+\Delta)})) < \tau^S_{min}$$

where

- f^U_{dist} : User profile distance function
- f^S_{sim} : Summary similarity function
- τ^U_{max} : Max. limit for two different user profiles to be mutually indistinguishable
- τ^S_{min} : Min. limit for two generated summaries w.r.t two different users to be mutually distinguishable
- τ^t_{max} : Max. time interval after which we compare the user profile to check drift in interests

$*$: $f^U_{dist}(\mathbf{u}_{(i,t)}, \mathbf{u}_{(i,t)}) = 0$ and $f^U_{dist}(\mathbf{u}_{(i,t)}, \mathbf{u}_{(i,t+\Delta)}) \in [0, 1]$

Example of how we can compare *Degree-of-Insensitivity* to Temporal Variance captured by different summarization models $M_{\boldsymbol{\theta}_x,\mathbf{u}}$, $M_{\boldsymbol{\theta}_y,\mathbf{u}}$ and $M_{\boldsymbol{\theta}_z,\mathbf{u}}$ is as follows:
If we have a metric that gives a score based on how insensitive the model is, i.e., how poorly a model generates a personalized summary, then as per figure 3, high score indicates poor personalized summary score given by metric, while low score indicates model generated better personalized summary for that news article. Expected e-DINS$_{TV}$ for a model $M_{\boldsymbol{\theta}_x,\mathbf{u}}$ is sample-average of its e-DINS$_{TV}$ score of all news over all user pairs.

B. Degree of Insensitivity w.r.t. Temporal Variance

$e-DINS_{TV}$ is a time sensitive measure which measures the insensitvity of a summarizer model towards the change in user profile over time. The change in user profile over time can be represented by the ***Angle of Deviation of user profile*** denoted by θ^U_i **??**(a). Similarly the ***Angle of Deviation of model generated summaries*** denoted by θ^S_i **??**(b)represents the shift in model generated summaries post the shift in user profiles.

These angles are calculated using the cosine rule. The side lengths represents the deviation of the User profile's distribution from the source document distribution and the deviation of the model generated summary's distribution from the source document distribution. This deviation is calculated using the Jensen-Shannon Divergence.

$\mathbf{u}_{(i,t)}$	$\mathbf{u}_{(i,t+\Delta)}$	D	$M_{\boldsymbol{\theta}_x,\mathbf{u}}$	$M_{\boldsymbol{\theta}_y,\mathbf{u}}$	$M_{\boldsymbol{\theta}_z,\mathbf{u}}$
			$D\text{-}Ins_{TVar(\boldsymbol{\theta}_x,\mathbf{u}_t,\mathbf{u}_{t+\Delta})}$	$D\text{-}Ins_{TVar(\boldsymbol{\theta}_x,\mathbf{u}_t,\mathbf{u}_{t+\Delta})}$	$D\text{-}Ins_{TVar(\boldsymbol{\theta}_x,\mathbf{u}_t,\mathbf{u}_{t+\Delta})}$
$\mathbf{Bob}_{13/01/23}$	$\mathbf{Bob}_{15/01/23}$	$\mathbf{News}_{13/01/23} \rightarrow \mathbf{News}_{13/01/23}$	0.59	0.17	**0.81**
$\mathbf{Alice}_{11/01/23}$	$\mathbf{Alice}_{17/01/23}$	$\mathbf{News}_{11/01/23} \rightarrow \mathbf{News}_{17/01/23}$	0.41	**0.74**	0.62

Table 2: Overall *Degree-of-Insensitivity* to Temporal Variance for a model $M_{\boldsymbol{\theta}_x,\mathbf{u}}$ is average of its $D_{KL_{(\boldsymbol{\theta}_x,\mathbf{u}_t,\mathbf{u}_{t+\Delta})}}$ score of all news over all user pairs

Fig. 3. Expected e-DINS$_{TV}$ for a model $M_{\boldsymbol{\theta}_x,\mathbf{u}}$ is sample-average of its e-DINS$_{TV}$ score of all news over all user pairs

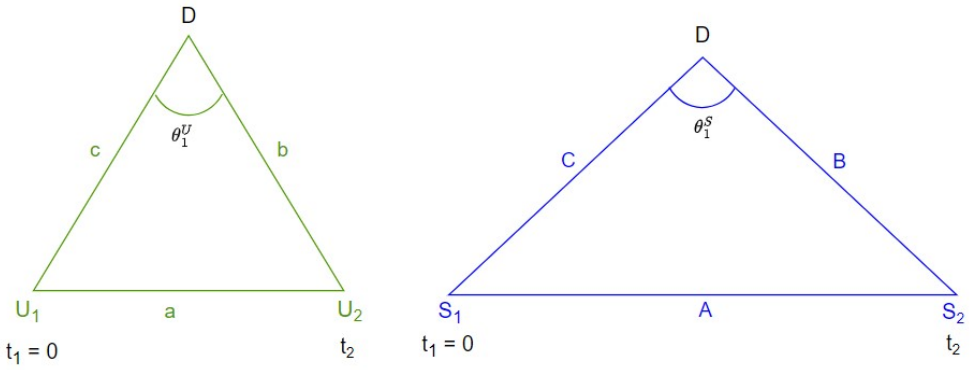

Fig. 4. (a)Angle of Deviation due to shift in user profile (b)Angle of Deviation due to shift in model generated summary

Using the cosine rule

$$a^2 = b^2 + c^2 - 2bc\cos\theta_1^U$$

$$\implies \theta_1^U = \arccos\frac{b^2 + c^2 - a^2}{2bc}$$

where
$a = JSD(U_1||U_2)$
$b = JSD(D||U_2)$
$c = JSD(D||U_1)$
Similarly we can calculate the Angle of deviation of model generated summaries

$$A^2 = B^2 + C^2 - 2BC\cos\theta_1^S$$

$$\implies \theta_1^S = \arccos\frac{B^2 + C^2 - A^2}{2BC}$$

where
$A = JSD(U_1||S_2)$
$B = JSD(D||S_2)$
$C = JSD(D||S_1)$

Since the change in user profile is not a function of time we cannot get a differentiable time dependent function of angular shift. But from the data we can get the user profiles at different time instants and thus from those shifts in user profiles the angular speed of the shift of user profile and the summarizer is calculated. (Figure 5)

$$\omega_1^U = \frac{\theta_1^U}{t_2 - t_1} \quad \omega_1^S = \frac{\theta_1^S}{t_2 - t_1}$$
$$\omega_2^U = \frac{\theta_2^U}{t_3 - t_2} \quad \omega_2^S = \frac{\theta_2^S}{t_3 - t_2}$$

Similarly

$$\omega_{n-1}^U = \frac{\theta_{n-1}^U}{t_n - t_{n-1}} \quad \omega_{n-1}^S = \frac{\theta_{n-1}^S}{t_n - t_{n-1}}$$

where n = number of versions of user profile and model generated summaries overtime

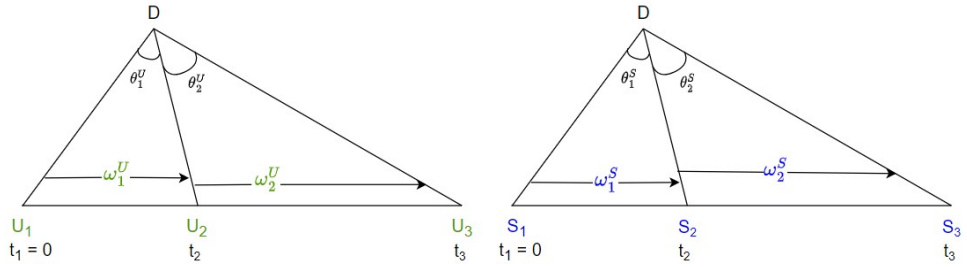

Fig. 5. (a)Angular speed of shift of user profile (b)Angular speed of shift of model generated summary

ω_i^U = angular speed of shift of the user profile in the time interval $[t_i, t_{i+1}]$
ω_i^S = angular speed of shift of the summaries in the time interval $[t_i, t_{i+1}]$

Now to measure the adaptivity of the model we need to analyze the rate of change of angular speed of the summarizer and user profile. A good summarizer should change the summaries according to the change in user profile along the same time instants. Thus the rate of change of the angular speeds helps us measure the rate at which the summarizer is able to change its model generated summaries to keep up with the change in user profile.

Now to measure the change in angular speeds

$$\alpha_1^U = \frac{\omega_2^U - \omega_1^U}{NetTime} \implies \alpha_1^U = \frac{\frac{\theta_2^U}{t_3 - t_2} - \frac{\theta_1^U}{t_3 - t_2}}{t_3 - t_1}$$

and

$$\alpha_1^S = \frac{\omega_2^S - \omega_1^S}{NetTime} \implies \alpha_1^S = \frac{\frac{\theta_2^S}{t_3 - t_2} - \frac{\theta_1^S}{t_3 - t_2}}{t_3 - t_1}$$

Similarly

$$\alpha_{n-2}^U = \frac{\omega_{n-1}^U - \omega_{n-2}^U}{NetTime} \implies \alpha_{n-2}^U = \frac{\frac{\theta_{n-1}^U}{t_{n-1} - t_{n-2}} - \frac{\theta_{n-2}^U}{t_n - t_{n-1}}}{t_n - t_{n-2}}$$

and

$$\alpha_{n-2}^S = \frac{\omega_{n-1}^S - \omega_{n-2}^S}{NetTime} \implies \alpha_{n-2}^S = \frac{\frac{\theta_{n-1}^S}{t_{n-1} - t_{n-2}} - \frac{\theta_{n-2}^S}{t_n - t_{n-1}}}{t_n - t_{n-2}}$$

Now to measure the insensitivity of the model towards the change in user profile overtime we check the difference of rate of change of angular speeds of the user profile and model generated summaries. The evaluation metric is the sum of the absolute values of these errors and normalized by their total number.

$$e_1 = |\alpha_1^U - \alpha_1^S|$$

$$e_2 = |\alpha_2^U - \alpha_2^S|$$

Similarly

$$e_{n-2} = |\alpha_{n-2}^U - \alpha_{n-2}^S|$$

The evaluation metric is the normalized sum of all the errors

$$e - D - INS_{TV} = \frac{e_1 + e_2 + ... + e_{n-2}}{n-2}$$

$$e\text{-}D\text{-}INS_{TV} = \frac{|\alpha_1^U - \alpha_1^S| + |\alpha_2^U - \alpha_2^S| + ... + |\alpha_{n-2}^U - \alpha_{n-2}^S|}{n-2}$$

$$\implies e - D - INS_{TV} = \frac{1}{n-2} \sum_{i=1}^{n-2} |\alpha_i^U - \alpha_i^S|$$

C. Shift in Source Document

Throughout the described procedure we have considered only one source document. In real life the source document also keeps changing. Consider a twitter thread; the user reads multiple tweets and his opinions and thoughts about the topic keeps on changing with each change in article or tweet. Thus the source Document also shifts. This change can also be incorporated in the metric. The figure 6 shows the shift of the source document.

The net angular shift of the user profile and the model generated summary changes by the shift in the source document. The new angular shift can be represented as

$$\theta_1^U = \theta_{11}^U + \theta_{12}^U \quad \theta_1^S = \theta_{11}^S + \theta_{12}^S$$

$$\theta_2^U = \theta_{21}^U + \theta_{22}^U \quad \theta_2^S = \theta_{21}^S + \theta_{22}^S$$

Similarly

$$\theta_n^U = \theta_{n1}^U + \theta_{n2}^U \quad \theta_n^S = \theta_{n1}^S + \theta_{n2}^S$$

V. Experimental Setup

A. Dataset

The requirements of the dataset for testing the evaluation metric includes the gold reference i.e. user written summaries from the source documents at different time points. There are 2 necessary conditions for the dataset to be eligible for temporal variance; Necessary condition 1: Each reader (human-judge) should be assigned at least two documents

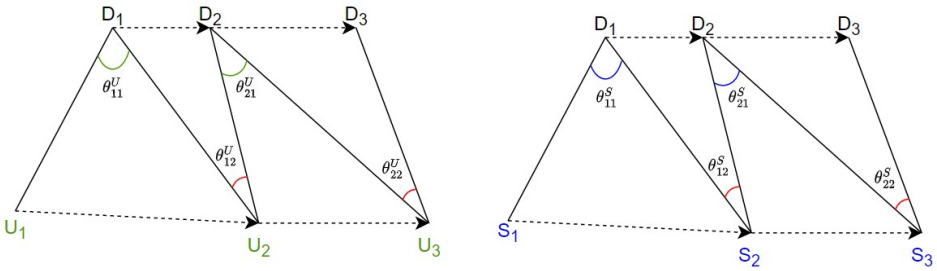

Fig. 6. Source Document Shift

Necessary condition 2: Each reader (human-judge) should be reading the assigned documents in a temporal sequence. However we can manipulate the datasets according to our requirements. Some of the datasets that we are considering are listed.

1) DUC (Document Understanding Conference) 2007 [5]: Documents summarized are from AQUAINT corpus comprising newswire articles from the Associated Press and New York Times (1998-2000) and Xinhua News Agency (1996-2000).
DUC 2007 consisted of 2 tasks:
Main Task: was real-world complex question answering, in which a question cannot be answered by simply stating a name, date, quantity, etc. Each topic and its document cluster were given to 4 different assessors (including the developer of the topic). The NIST assessors created a 250-word summary of the document cluster that satisfies the information need expressed in the topic statement i.e. that answers the questions in the topic statement. Test Data contains 10 topics.

Update Task: The task was to produce short (100 words) multi document update summaries based on previously read articles. For each topic the documents were ordered chronologically and then partitioned into A-C, where the timestamps on all the documents in each set are ordered such that time(A) $<$ time(B) $<$ time(C). A summary of documents in cluster A. An update summary of documents in B. An update summary of documents in C. The purpose of each update summary was to inform the reader of new information about a particular topic. Test data consists of 10 topics 25 documents per topic, 10 documents in set A, 8 in Set B and 7 in Set C.

2) TREC 2015 Temporal Summarization Task [6]: The aim of this task is to emit series of sentence updates over time about a named event, given a high volume stream of input documents. Temporal summarization task focuses on large events with a high impact, such as protests, accidents or natural disasters. Each event is represented by a topic description, providing a textual query representing that event, along with start and end timestamps defining a period of time within which to track that event. (figure 7)

```
<event>
  <id>1</id>
  <title>2012 Buenos Aires rail disaster</title>
  <description>...</description>
  <start>1329910380</start>
  <end>1330774380</end>
  <query>buenos aires train crash</query>
  <type>accident</type>
</event>
```

Fig. 7. TREC Temporal Summarization Task

3) TAC 2008: The TAC 2008 update summarization task provides test data which consists of 48 topics; each topic containing 20 documents.

For content evaluation 4 model pyramid model was used and was performed by NIST assessors.Parameters were the reference topic-specific multi-document summary, set of SCUs generation per topic and Mapping reference multi document summary to SCUs. Readability/Fluency and Responsiveness was also considered.
Pyramid model (Figure 8) doesn't capture subjectivity. Rather it tries to objectify(generalize) different summaries by creating SCUs (summary content units). Further these SCUs used to evaluate system summary.

4) TES 2012-2016 [7]: The Twitter events dataset from 2012 to 2016 was used to extract tweets using the provided set of IDs. The dataset contains approximately 150 million tweet IDs, which were obtained by crawling hashtags and keywords related to 30 different events. In order to obtain relevant summaries, the existing summaries on the Wikipedia Current Events Portal (WCEP) were manually extracted for each event, covering the same time period as the tweets. Since WCEP presents summaries on a daily basis, the time increment for each summary is one day.

Figure 9 contains some statistics regarding the TES 2012-2016 collection. Out of the 30 events, two events, namely the SXSW festival in 2012 and the St. Patrick's Day 2014 celebrations, were not mentioned in WCEP and were therefore not included in the dataset. The remaining events had between

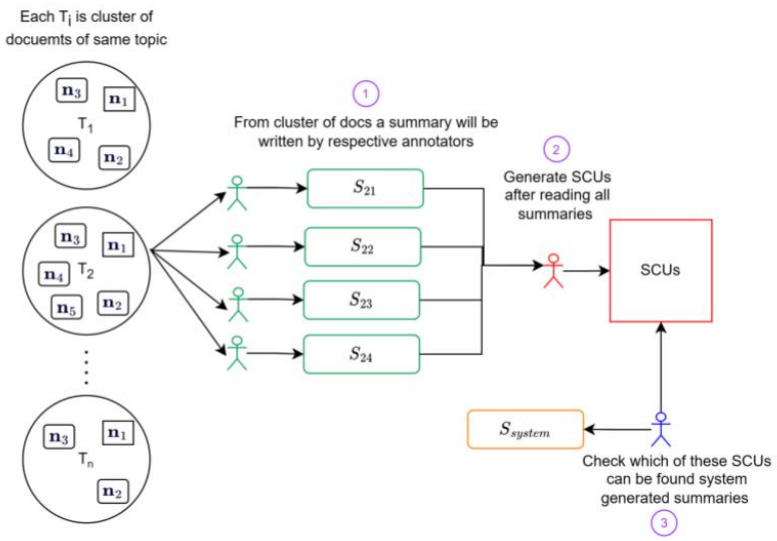

Fig. 8. Pyramid Model

Event	Number of ids	Number/% of retrieved tweets	Number of subevents in WCEP
euro 2012	8,992,157	5,625,286 / 63	52
hurricane sandy 2012	14,914,566	8,399,189 / 56	9
mexican election 2012	191,788	125,093 / 65	1
obama romney 2012	10,146,517	5,033,962 / 50	8
superbowl 2012	1,659,475	1,027,944 / 62	2
us election 2012	1,740,258	1,001,114 / 58	8
boston marathon bombing 2013	3,430,387	1,738,635 / 51	4
ebola 2014	986,525	663,081 / 67	8
ferguson 2014	8,782,071	5,105,294 / 58	7
gaza under attack 2014	2,886,322	1,407,497 / 49	17
hongkong protests 2014	1,188,372	764,371 / 64	11
indyref 2014	1,524,166	952,495 / 62	5
ottawa shooting 2014	1,075,864	675,022 / 63	1
sydney siege 2014	2,157,879	1,246,560 / 58	3
typhoon hagupit 2014	264,626	161,666 / 61	4
charlie hebdo 2015	18,940,619	11,101,149 / 59	10
germanwings crash 2015	2,648,983	1,583,392 / 60	8
hurricane patricia 2015	1,151,220	578,559 / 50	1
nepal earthquake 2015	12,004,187	7,364,891 / 61	19
paris attacks 2015	29,821,274	16,617,011 / 56	32
refugees welcome 2015	1,743,153	1,064,537 / 61	81
brexit 2016	1,826,290	1,001,836 / 55	3
brussels airport explosion 2016	5,869,990	3,328,377 / 57	10
hijacked plane cyprus 2016	702,586	424,930 / 60	3
irish general election 2016	758,803	541,989 / 71	4
lahore blast 2016	1,149,253	685,922 / 60	2
panama papers 2016	5,044,379	3,368,653 / 67	19
sismo ecuador 2016	1,007,867	727,112 / 72	9
Total	142,609,577	82,315,567 / 58	341

Fig. 9. Some statistics for the TES 2012-2016 collection.

1 to 81 sub-events, with an average of 12 sub-events per event. Some time windows had multiple sub-events, and some sub-events contained a wealth of facts, resulting in gold standard time windows consisting of multiple sentences. The dataset

consisted of 82,315,567 tweets, which were retrieved from the original Twitter events 2012-2016 collection, comprising only 58% of the released IDs, with a minimum of 49% and a maximum of 72% for a single event. The smallest number of retrieved tweets for an event was 125,093 for the Mexican election in 2012, and the largest was 16,617,011 for the Paris attacks in 2015. On average, the TES 2012-2016 dataset comprised 2,939,842 tweets per event. Additionally, two different Oracle summaries were constructed using two different relevance metrics, namely ROUGE-2 F and Cosine measures, to determine the upper bounds of summarization methods.

5) PENS: We are using the test set of PENS(Personalized News headlineS) dataset [4]. Format of dataset is as per given in figure 11. Headlines can be considered as TLDR summary. Click behaviors of 103 english native speakers are collected and more than 20,000 manually written personalized headlines of news articles, regarded as the gold standard of user-preferred titles are included in the dataset. Test set created in 2 stages: In the first stage, Each user reads 1000 news articles and selects at least 50 articles in which he is interested. Headlines are randomly selected and arranged by their first exposure time. In the second stage, each person writes their preferred headlines for 200 articles without knowing the original news title. These news articles are excluded from the first stage.

Consider the example shown in fig. 10, in which underlined words and colored words represent the correlated words in the manually-written headlines, clicked news, and the generated headlines, respectively.

Case 1. Original Headline:	Venezuelans rush to Peru before new requirements take effect
Pointer-Gen:	Venezuelans rush to Peru
user A written headline:	New **requirements** set to take effect causes **Venezuelans** to rush to **Peru**
NAML+HG for user A:	**Peru** has stricter entry **requirements** for escaping **Venezuelans** on that influx.
Clicked News of user A:	1. Peru and Venezuela fans react after match ends in a draw 2. Uruguay v. Peru, Copa America and Gold Cup, Game threads and how to watch
user B written headline:	**Venezuelan migrants** to Peru face danger and discrimination
NAML+HG for user B:	Stricter entry requirements on **Venezuelan migrants** and **refugees**.
Clicked News of user B:	1. Countries Accepting The Most Refugees (And Where They're Coming From) 2. Venezuelan mothers, children in tow, rush to migrate

Fig. 10. Example of personalized headline

Column	Example Context	Description
userid	NT1	The unique ID of 103 users
clicknewsID	N108480,N38238,N35068, ...	The user's historical clicked news collected at the first stage
posnewID	N24110,N62769,N36186, ...	The exhibited news for each user at the second stage
rewrite_titles	"Legal battle looms over Trump EPA\'s rule change of Obama\'s Clean Power Plan rule ...	The manually-written news headlines for the exhibited news articles and can be split by '#TAB#'

Fig. 11. PENS test set format

Conclusion

In conclusion, the development of an evaluation metric to evaluate personalized summarizers based on temporal variance has been successfully achieved in this semester project. The project aimed to address the limitations of traditional summarization evaluation metrics by considering the temporal variance of the information in the source text. The evaluation metric proposed in this project takes into account the relevance of the summary to the source text and its temporal variance, which reflects the importance of up-to-date information in the summary. The results of this project can be used to guide the development of future personalized summarization models, and they can also be used to inform the evaluation of existing models.

Future Work

The Future work of this project includes finding an appropriate dataset and manipulating it according to the needs of the formula to get the gold reference and model generated summaries and testing the proposed evaluation metric on different models. The correlation of the proposed evaluation metric also needs to be found.

Acknowledgment

I would like to express my sincere gratitude to Professor Sourish Dasgupta for guiding me throughout the course of this project. His insightful comments, valuable suggestions, and constant encouragement have been instrumental in the successful completion of this project. I am deeply indebted to him/her for his/her support and guidance.

I would also like to thank Rahul Vansh, an MTech student, for his help and support during the project. His technical expertise, cooperation, and willingness to assist have been of great help to me, and I cannot thank him/her enough for his/her contribution to this project.

References

[1] D. O. Cajueiro, A. G. Nery, I. Tavares, *et al.*, *A comprehensive review of automatic text summarization techniques: Method, data, evaluation and coding*, 2023. arXiv: 2301.03403 [cs.CL].

[2] S. Ghodratnama, M. Zakershahrak, and F. Sobhanmanesh, "Adaptive summaries: A personalized concept-based summarization approach by learning from users' feedback," *CoRR*, vol. abs/2012.13387, 2020. arXiv: 2012.13387. [Online]. Available: https://arxiv.org/abs/2012.13387.

[3] S. Ghodratnama, A. Beheshti, M. Zakershahrak, and F. Sobhanmanesh, "Extractive document summarization based on dynamic feature space mapping," *IEEE Access*, vol. 8, pp. 139 084–139 095, 2020. DOI: 10.1109/ACCESS.2020.3012539.

[4] X. Ao, X. Wang, L. Luo, Y. Qiao, Q. He, and X. Xie, "PENS: A dataset and generic framework for personalized news headline generation," 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)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 82–92. DOI: 10.18653/v1/2021.acl-long.7. [Online]. Available: https://aclanthology.org/2021.acl-long.7.

[5] R. Witte, R. Krestel, and S. Bergler, "Generating update summaries for duc 2007," in *Proceedings of the Document Understanding Conference*, 2007, pp. 1–5.

[6] P. Wang and W. Li, "Iscasir at trec 2015 temporal summarization track.," in *TREC*, 2015.

[7] A. Dusart, K. Pinel-Sauvagnat, and G. Hubert, "Tssubert: How to sum up multiple years of reading in a few tweets," *ACM Transactions on Information Systems*, 2023.