IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 January 2023, accepted 13 February 2023, date of publication 23 February 2023, date of current version 28 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3247948

RESEARCH ARTICLE

Aligning Comments to News Articles on a Budget

JUMANAH ALSHEHRI, MARTIN PAVLOVSKI, EDUARD DRAGUT, (Member, IEEE), AND ZORAN OBRADOVIC, (Senior Member, IEEE)
Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA 19122, USA

Corresponding author: Zoran Obradovic (zoran.obradovic@temple.edu)

This work is supported in part by the U.S. NSF awards 2026513 and 1838145, and the U.S. Army Research Laboratory subaward 555080-78055 under Prime Contract No.W911NF2220001, the U.S. Army Corp of Engineers Engineer Research and Development Center under Cooperative Agreement W9132V-22-2-0001, and Temple University office of the Vice President for Research 2022 Catalytic Collaborative Research Initiative Program AI & ML Focus Area. In addition, this research includes calculations carried out on HPC resources supported in part by the U.S. NSF through major research instrumentation grant number 1625061 and by the U.S. Army Research Laboratory under contract number W911NF-16-2-0189. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

ABSTRACT Disagreement among text annotators as a part of a human (expert) labeling process produces noisy labels, which affect the performance of supervised learning algorithms for natural language processing. Using only high agreement annotations introduces another challenge: the data imbalance problem. We study this challenge within the problem of relating user comments to the content of a news article. We show that traditional techniques for learning from imbalanced data, such as oversampling, using weighted loss functions, or assigning weak labels using crowdsourcing, may not be sufficient for modeling complex temporal relationships between news articles and user comments. In this study, we propose a framework for aligning comments and articles 1) from *imbalanced news data* characterized with 2) different degrees of *annotator agreement*, under 3) a *constrained budget* for human labeling and computing resources. Within the framework, we propose a Semi-Automatic Labeling solution based on Human-AI collaboration. We compare our proposed technique with traditional data imbalance handling techniques and synthetic data generation on the *article-comment alignment problem*, where the goal is to determine a category of an article-comment pair that represents how relevant the comment is to the article. Finding an effective and efficient solution is essential because it is time-consuming and prohibitively costly to manually label a sufficiently large amount of article-comment pairs based on the semantic understanding of an article and its comments. We discover that the Human-AI collaboration outperforms all alternative techniques by 17% of article-comment alignment accuracy. When there is no time or budget for re-labeling some article-comment pairs, we found that synonym augmentation is a reasonable alternative. We also provide a detailed analysis of the effect of humans in the loop and the use of unlabeled data.

INDEX TERMS Annotators' disagreement, article-comment alignment, imbalance classes, multi-class classification.

I. INTRODUCTION

The underlying assumption in supervised learning is that ground truth, which humans usually obtain, is error-free. In practice, this assumption is not always true; different tasks have different levels of difficulty, which for certain applications makes it challenging for humans to agree on a ground truth [28]. In many practical problems, humans do not agree on their annotations [4], providing noisy labels that affect the downstream performance of machine learning models [8]. One such problem is the Article-Comment Alignment Problem (ACAP) [2], where the goal is to classify an article-comment pair into one of several categorical relevancy levels. For example, an annotator might consider article-comment pairs relevant, while another might consider them irrelevant. Such disagreement may be due to factors outside reading comprehension (e.g., background knowledge). For example, consider an article on the ongoing war in Ukraine and a user comment on that article that talks about the war in Kosovo. One annotator may be aware of that event, while another may

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

not be aware of that event, which happened in 1999. They will likely give different annotations for the relationship between the article and the comment. We found that focusing on data points with high agreement and ignoring data points with low agreement leads to class imbalance issues. This typically occurs when the examples with high disagreement belong to one class, which is the focus of our study.

Data imbalance is characterized by a skewed distribution of classes [22]. More precisely, such imbalance arises when a class or some classes are not sufficiently represented in the training examples compared to other classes. Class imbalance is a property of real-life data and a well-studied machine-learning problem. A line of research focuses on using evaluation measures that are more appropriate for imbalanced data, such as the area under ROC curve [13]. Other research aims at teaching a model to weigh instances according to their frequency across classes. For example, an error on instances from the minority class may be penalized more during the training process compared to the majority class [6]. This action will force the model to pay more attention to the minority class examples [14]. Some of the traditional approaches focus on balancing the number of examples using oversampling and undersampling techniques [29], and some research focuses on developing toolkits that can serve this purpose, such as [33]. Oversampling and undersampling cannot be applied to all problems, particularly not in the presence of noisy labels that may further exacerbate the classification performance and distort a classifier's decision boundary. Other research proposes generating synthetic examples to augment the minority class, such as Synthetic Minority Oversampling Technique (SMOTE) [1].

This work proposes a framework to select the most effective and efficient technique given a constrained budget and desired performance. Hence, we consider Human-AI collaboration when such teaming is feasible and compare it to a fully automatic technique requiring minimal human-based annotation. As a case study, we tackle the class imbalance issue in ACAP, which we describe in Section II. This problem is challenging to address using crowdsourcing since reading articles and their associated comments is burdensome and time-consuming. In addition, this requires a semantic understating of the given input to grant the most appropriate label that reflects the level of relevancy between an article and its corresponding comments. Therefore, we omit to utilize crowdsourcing to address the problem considered in this work and instead aim at providing possible answers to the following research questions:

- Q1: Does the proposed human-AI technique enhance the model performance compared to baselines?
- Q2: To what extent does the human-in-the-loop approach and the amount of unlabeled data affect a Human-AI team's performance?
- Q3: How does synthetic data generation affect the article-comment alignment performance?
- Q4: What are each technique's marginal benefit and marginal cost?

II. PROBLEM FORMULATION

A. ARTICLE-COMMENT ALIGNMENT PROBLEM (ACAP)

Assume a source (e.g., a news outlet) $s \in S$ that consists of n articles and their comments $s = \{\langle a_1, (c_{11}, .., c_{1m_1})\rangle, .., \langle a_n, (c_{n1}, .., c_{nm_n})\rangle\}$, where the article a_i is associated with m_i comments, for each $i = 1, \ldots, n$. ACAP is the task of finding a function $f(a_i, c_{ij})$ that maps an article-comment pair to a target Y, which is a class that indicates the relevance level between a given comment and its corresponding news article. In this study, four ordinal classes are considered. The classes are *1) Relevant* - the comment's content discusses the same matter as the article. *2) Same Category* - the comment is not relevant; however, it discusses an issue from the same category as the article. For example, an article that discusses the 2022 Russian invasion of Ukraine and a comment discussing the Iraq invasion of Kuwait in 1990. In this case, the article and comment do not discuss the same issue but are within the same category (politics). *3) Same Entities* - the comment is not directly relevant, however, it mentions the same entities appearing in the article but in a different scope. For example, consider an article discussing the Russian invasion of Ukraine and mention the Russian President *Vladimir Putin*. In contrast, a comment may discuss the UK's decision not to invite *Putin* to Queen Elizabeth's funeral. Here, both the article and the comment mention Putin but in a different context. Finally, *4) Irrelevant* - the comment content is irrelevant to the article and does not belong to any of the above classes. Our classes are ordinal with some overlap; the more a comment context differs from the article, the higher the distance between the article and the comment.

B. CHALLENGES

ACAP requires a comprehension of both the article and its comments, making it challenging and time-consuming. To label the data, we provide annotators with instructions that we discuss in Section V-A. Later, we interviewed annotators to get their feedback regarding the labeling process. We found that they had difficulty distinguishing between the *Relevant* class and *Same Category* class, resulting in a high disagreement between the annotators. According to the Fleiss Kappa statistic, the level of disagreement between annotators is mostly fair. Therefore, we define two types of data points in each source: Gold Standard (GS) and Noisy Examples (NE). The GS examples are examples for which the annotators predominantly agree, while NE examples represent those examples for which annotators disagree with each other for the most part. In other words, GS are examples with low standard deviation (σ) of annotation disagreement, computed as

$$\sigma = \sqrt{\frac{\Sigma|x - \bar{x}|}{n}}, \tag{1}$$

where n is the total number of annotators, x is the class given by the n^{th} annotator, and $\bar{x}$ is the mean of annotations given by all annotators. In this study, we used three annotators and $\sigma = 0.5$ to divide the article-comments pairs into GS and NE

examples. If all annotators agree or one annotator disagrees with a difference of one (e.g., the annotator labels a comment as *Same Category*, i.e. label 2, while all other annotators label it as *Same Entities*, i.e. label 3), using $\sigma = 0.5$ will not change the aggregated (final) class obtained by an aggregating schema. On the other hand, $\sigma > 0.5$ would mean that all three annotators disagree, or at least one of them disagrees with a larger difference (2 or 3), which affects the aggregated label more.

III. RELATED WORK

In supervised learning, challenging problems [28] are subject to annotators' disagreement [7] between humans. A line of research [4], [8], [12] shows that the presence of noisy labels is highly correlated with the performance of a classification model [8]. Focusing only on examples with high agreement might lead to class imbalance [14], [22]. Research focuses on addressing the data imbalance problem from an evaluation perspective [13], and model point of view [3]. In contrast, other works argue that producing weak labels obtained by active learning [5] and crowdsourcing [35] will overcome the class balancing issue. However, sometimes crowdsourcing might not be a suitable approach, considering that resources (i.e., budget) are limited in many large-scale applications. Therefore, another approach is to utilize unlabeled data to produce pseudo-labels using semi-supervised techniques [20].

In this work, we focus on solving the imbalance issue by generating synthetic data and leveraging many unlabeled data. The text requires systematic techniques to generate synthetic data that are from the same distribution as real data. Some research focuses on generating synthetic examples using summarization techniques [23], [26], [27], and synthetic data generation techniques such as synonym augmentation [18]. Another line of work is focused on generating new labeled examples by leveraging unlabeled data [10], [34].

We consider two techniques to enhance our data. First, we utilize summarization and synonym augmentation to generate synthetic data. Second, we utilize unlabeled data to generate more labeled examples. We make the following contributions in this work: 1) utilize summarization, and synonym augmentation to ACAP [16], [32]; 2) develop a framework for leveraging unlabeled data with the human-in-the-loop approach; and 3) propose a utility function that allows practitioners to select an appropriate approach based on their budget and desired performance.

IV. THE PROPOSED TECHNIQUE

Here we define our proposed Semi-Automatic Labeling technique for article-comment alignment with a high imbalance of classes due to high human-to-human labeling disagreement. We propose to utilize unlabeled data using a semi-automatic labeling method with Human-AI collaboration.

Semi-Automatic Labeling focuses on utilizing unlabeled data through semi-supervised techniques since unlabeled data is cheaper to obtain. One of the most widely used

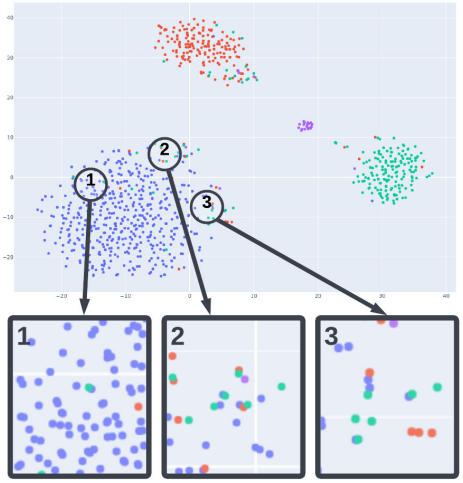


FIGURE 1. An example of comment embedding visualization for GS Fox News using t-SNE. Each color represents a class. A zoomed-in case 1) shows misclassified examples, while cases 2) and 3) show misclassified examples with high overlap between different classes, making it more difficult to assign the aggregated class for these examples.

semi-supervised methods is pseudo labeling [20], which creates weak labels with the use of Long-Short Term Memory (LSTM) [15] networks. We use a different approach, relying on a Human-AI team. Here we start by learning the distribution of comments in the GS data using the pre-trained BERT [9] and visualizing the distribution in two dimensions using t-distributed stochastic neighbor embedding (t-SNE) [24]. Figure 1 shows an example of comment embedding visualization for a particular news source (Fox News) using t-SNE. From the visualization, we identify misclassified comments that do not belong to the same cluster based on its distribution and ask humans to relabel them by reading the article and the comment and checking if the given class is correct or not. This process is repeated until we reach an error rate less than λ, where λ is a hyperparameter. A larger λ means that we relax the error rate constraint, resulting in fewer iterations and fewer contributions from the human side and vice versa. To optimize λ, we run multiple experiments with different values of λ; we found that $\lambda = 0.05$, allowing a 5% error rate in our GS examples, results in a satisfactory performance with a reasonable number of human correction iterations. Finally, we learn the embedding of the unlabeled data and label them according to their nearest neighbors using the k-nearest neighbors (KNN) algorithm. We tested different values of $k = 2, 3, 5$ and found that $k = 3$ is the most suitable option. Note that unlabeled data is not part of the evaluation and testing sets during training and evaluation. Algorithm 1 shows the process of the Semi-Automatic Labeling technique. The inputs to the algorithm are the GS examples, unlabeled examples, and λ, which is the allowed error rate given by the user. The remaining hyperparameters in the algorithm are described as follows: 1) c parameter is a counter for the number of rounds the human has to perform certain corrections. 2) $\mathbb{E}$ represents misclassified examples, which is a subset of GS. We identify

these instances by visualizing the comment distribution using t-SNE. 3) ϵ actual error rate, calculated as $count(\mathbb{E})$, the number of misclassified examples divided by $count(GS)$, the total number of examples in GS. 4) $\mathbb{E}_{new}$ represents the examples that humans correct. 5) GS_{update} is the merge of GS examples after we remove the misclassified examples and the corrected examples $\mathbb{E}_{new}$. After c counts, where $\epsilon \leq \lambda$, we start the prediction process for unlabeled examples as follows. 1) Train a *KNN* model using GS_{new} examples. 2) Learn the embedding of the unlabeled data using $\hbar$, the last hidden state of *BERT* that is pre-trained on GS. 3) Predict the labels for unlabeled data using the trained *KNN* model based on their three nearest neighbors. The outputs of the algorithm are GS_{update}, which is the final updated *GS* after the human correction, $unlabeled_{new}$, the new labeled data, and c, which represents the total number of rounds that the human has to perform some corrections.

Algorithm 1 Semi-Automated Labeling

Procedure: $SemiAutomated(GS, unlabeled, \lambda)$

```
/* counter for human correction
   rounds                                  */
c = 0
/* While actual error rate (ϵ) is
   greater than allowed error rate
   (λ)                                     */
while (ϵ > λ) do
    /* Learn GS comments' embeddings
       */
    emb_GS = BERT(GS)
    /* Identify misclassified
       examples                            */
    E = tSNE(emb_GS)
    /* update the error rate (ϵ)          */
    ϵ = count(E) / count(GS)
    /* Apply human correction to E        */
    E_new = HumanCorrection(E)
    GS_update = Merge(GS, E_new)
    c++
/* Learn unlabeled embedding using
   the pre-trained BERT's last
   hidden state (ħ)                        */
emb_unlabeled = ħ(unlabeled)
KNN_train = KNN(GS_update)
unlabeled_new = KNN_train(emb_unlabeled)
```

Output: GS_{update}, $unlabeled_{new}$, c

V. EXPERIMENTS

This section describes the data, classification model, comparison techniques, and experimental settings.

A. DATA AND LABELING

We collected news articles and comments between 2015 and 2017 from multiple news sources. We chose five news sources representing the ACAP. The data contains articles and comments with a broad range of lengths and a different number of comments per article. Three English speakers generated labels to annotate the article-comment pairs [2]. The annotators manually and independently mapped the pairs to one of the four proposed classes, 'Relevant', 'Same Entities', 'Same Category', and 'Irrelevant'. We provide the annotators with the following: 1) article-comment pairs without the surrounding context (i.e., the parent and child comments), and 2) the four classes of relevance-level categories with an explanation and an example for each. We assign the final label of each pair using an averaging aggregation scheme. We study ACAP on five news sources where each source contains news articles and comments from Wall Street Journals (WSJ), Fox News (FN), Daily Mail (DM), The Guardian (TG), and Market Watch (MW), where each source consists of 1,000 article-comment pairs. Table 1, shows the descriptive statistics of each source. We can see that the relevant class is the most underrepresented in most sources. By performing a Fleiss Kappa agreement analysis on GS examples, we observe that the agreement score improved between $[12\% - 22\%]$, where the actual scores are between $[0.44 - 0.60]$, and the agreement scores for all sources shift from the 'fair' range to the 'moderate' range. However, even when using GS examples, the agreement did not reach a substantial or almost perfect agreement based on the interpretation of the Kappa score. Because for most pairs in the GS set, one annotator assigns a label with a distance of 1 compared to the other annotators. For instance, two annotators assign a class value of 1 (relevant), while one annotator assigns a class value of 2 (same category). In this work, labeled NE examples are treated as unlabeled examples.

TABLE 1. Sources' statistics. # tokens (art.) represent the articles' maximum sequence length expressed as the number of words after text preprocessing. # tokens (comm.) represent the comments' maximum sequence length expressed as the number of words after text preprocessing. GS are gold standard examples (article-comment pairs), NE stands for noisy examples (article-comment pairs) and Unlabeled are all unlabeled examples (article-comment pairs). The last four rows represent the proportion of each class in GS (article-comment pairs).

Source	WSJ	FN	DM	TG	MW
# tokens (art.)	128	256	384	512	512
# tokens (comm.)	64	32	32	64	32
# GS (pairs)	543	876	833	795	862
# NE (pairs)	457	124	176	205	138
# Unlabeled (pairs)	$3.8K$	$18.6K$	$30K$	$30K$	$4.4K$
% Class Relevant	10.3%	3.0%	18.0%	9.3%	4.8%
% Class Same Ent.	19.3%	20.4%	47.6%	50.4%	38.1%
% Class Same Cat.	16.7%	21.5%	14.5%	12.5%	28.6%
% Class Irrelevant	53.5%	54.9%	19.8%	27.6%	28.3%

B. DOWNSTREAM CLASSIFICATION MODEL

To classify article-comment pairs, we use BERTAC [2], which utilizes $BERT_{base}$ architecture to predict the article-comment pair class. BERTAC allows us to learn more expressive embeddings of articles and comments. To address ACAP, we combine an article and its comment into a pair of segments. We aim to use BERT's self-attention mechanism

and bidirectional cross-attention in an end-to-end manner to encode the relevance between an article comment pair.

C. BASELINE TECHNIQUES

We compare the proposed technique to the original labels obtained by annotators and GS. In addition, traditional data imbalance method and generating synthetic comments.

1) ORIGINAL

The initial data provided to annotators for labeling consists of 1, 000 article-comment pairs for each source. These data contains GS and NE data points, where GS examples represent $54\% - 87\%$ of the data, while NE represents $13\% - 46\%$.

2) GS

Only labeled data points with high agreement scores (according to the definition described in Section II) are used for learning. Table 1 shows the number of data points in GS and the proportions of each class in percentages. Note that in all data, the 'Relevant' class represents the minority class with around $29\% - 51\%$ fewer data points than the majority class, which is 'Irrelevant' in all applications except for FN and DM, where the majority class is 'Same Entities'.

3) TRADITIONAL DATA IMBALANCE TECHNIQUES

a: OVERSAMPLING

The goal of oversampling is to distribute the classes uniformly. This technique expands the minority classes by duplicating data points from the minority class. Here we utilize Random Oversampling [25] by selecting samples at random with replacement. We choose to oversample all classes to uniform the number of data points in each class. Using this method, we generate $250 - 390$ data points per class.

b: WEIGHTED LOSS

In this technique, we modify the model loss function to account for the minority class more by assigning a higher weight to examples from the minority class during the training process. To achieve this, we assigned a weight to each example in the data, where the weight is calculated as follows,

$$w = \begin{cases} 1 & \text{if } D \notin Class_{min}, \\ \dfrac{count(D_{max})}{count(D_{min})} & \text{otherwise.} \end{cases} \tag{2}$$

The example weight is equal to 1 if the example does not belong to the minority class D_{min}, which means we are using the model with the original loss function. In contrast, the example weight is the total number of the majority class examples $count(D_{max})$ divided by the total number of the minority class examples $count(D_{min})$. D denotes the sample of the data that belongs to one class.

4) GENERATING SYNTHETIC DATA

In this approach, we utilize our understanding of the problem that the relevant class is underrepresented in our data, as shown in Table1. This technique aims to produce synthetic relevant comments to augment the dataset. Hence, we produce synthetic comments using 1) Extractive Summarization and 2) Synonym Augmentation. To ensure that none of the synthetic data is leaked into the testing set, in all our experiments, we split the GS data into training, validation, and testing datasets, and then we merge the generated data with the training set.

a: EXTRACTIVE SUMMARIZATION

This technique utilizes the articles to generate relevant comments by selecting sentences in the article that are highly representative of the full article and treating them as comments. We utilize statistical methods [26] and sentence similarity based on graphs techniques [27] to generate new relevant comments. We extract sentences containing words with higher frequencies, such as TF-IDF. These sentences are considered the most representative of the article. On the other hand, we generate additional comments by converting the article and each sentence of the article to a matrix and then calculating the similarity between the sentence matrix and article matrix using Euclidean distance. Then we employ PageRank [30] to rank sentences based on their similarity. Using both techniques helps produce synthetic yet relevant comments from the same distribution as the articles. Note that in this work, we do not intend to propose a new text summarization technique; therefore, we utilize more reliable and well-known methods.

b: SYNONYM AUGMENTATION

Using relevant comments in our GS data, we generate new comments by replacing some words with their synonyms [18]. Here we utilize Global Vectors for Word Representation (GLOVE) [31] to substitute the verbs and nouns with their synonyms. This technique helps produce synthetic relevant comments from the same distribution as the original comments generated by users.

c: COMBINED SYNTHETIC GENERATION

To understand the effect of each synthetic data generation technique, we also compare to using merged synthetic examples produced by Extractive Summarization and Synonym Augmentation.

D. UTILITY FUNCTION

This function aims to provide an approximation of the cost of resources to aid the decision in terms of which approach to take. To that end, we define a utility function that penalizes the performance of a given approach with its associated cost and thus can be used to find the most effective and efficient approach among multiple approaches. The utility function is calculated as follows,

$$f^* = \arg\max_{f_k} \mathbb{U}(f_k), \quad k = 1, \ldots, K, \tag{3}$$

Here, $\mathbb{U}(f_k)$ is the performance of the k-th approach penalized by its associated cost, defined as:

$$\mathbb{U}(f_k) = |p_k - (p_k^{-1} \cdot \acute{c_k})|, \tag{4}$$

where p_k is a performance measure (in this study we measure performance in terms of predictive accuracy); p^{-1} is the inverse value of p which is used to prevent $\mathbb{U}(f_k)$ from vanishing, $\acute{c_k}$ is the normalized value of c_k defined as the cost of approach k. We normalize the value of c_k to make sure that both the performance and cost values fall within the same range of [0, 1]. The normalized cost $\acute{c_k}$ is calculated as follows,

$$\acute{c_k} = \frac{c_k - min(\vec{c})}{max(\vec{c}) - min(\vec{c})}, \tag{5}$$

where $\vec{c} = [c_1, \ldots, c_K]$ is a vector that contains the costs of all approaches. The cost of the k-th approach is defined as

$$c_k = (t_h \cdot \delta_h) + (t_{cpu} \cdot \delta_{cpu}), \tag{6}$$

where t_h and δ_h denote the time allocated for humans in the loop and the human cost per unit (hour), respectively; while t_{cpu} is the CPU time and δ_{cpu} is the CPU cost per time unit. In this study, since humans participated voluntarily, $\delta_h = \delta_{cpu} = 1$. However, the cost is still accounted for in the equation for reproducibility and generalization purposes.

VI. RESULTS AND DISCUSSIONS

To answer our research questions, we start by 1) comparing the overall performance between the baselines and proposed approach, followed by a detailed analysis of the performance of predicting each class separately. 2) Show the effect of humans and the amount of unlabeled data in the Human-AI team technique. 3) Analyze the impact of different synthetic data generation techniques. 4) We present the utility function analysis that aims to determine the best approach.

A. Q1: PERFORMANCE OF BASELINES AND THE PROPOSED TECHNIQUE

1) OVERALL PERFORMANCE

Here we compare the performance of the baseline techniques with the proposed techniques. As shown in Table 2, Semi-Automatic Labeling, where we utilize unlabeled data in a Human-AI team, clearly outperforms all other proposed and baseline techniques. Semi-Automatic Labeling increases accuracy between $14\% - 17\%$ compared to Original and GS. By comparing synthetic data generation techniques, we see that Synonym Augmentation is slightly better than Extractive Summarization with an increase in accuracy between $1\% - 8\%$. We believe that the reason behind this is the ability of Extractive Summarization to generate comments based on the article. In contrast, Synonym Augmentation generates comments based on the original comments from the same distribution as that of a given article; this shift in distribution leads to an increase in performance. Furthermore, looking at the model stability, based on the standard deviation in classification accuracy, we observe that the proposed techniques are more stable than other techniques, particularly compared to oversampling and weighted loss, where the standard deviation is at least five times higher. Finally, we conclude that traditional class imbalance techniques, such as Oversampling and Weighted Loss, do not answer the ACAP class imbalance problem.

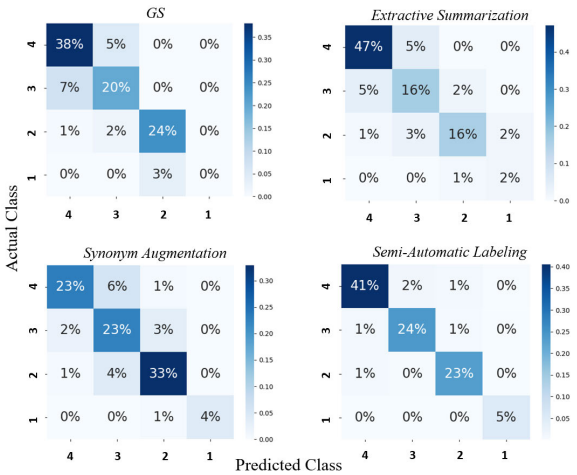


FIGURE 2. FN confusion matrices on the test set for GS and proposed techniques. Classes are as follows: 4='Irrelevant', 3='Same Entities', 2='Same Category', and 1='Relevant'.

2) WITHIN-CLASS PREDICTION PERFORMANCE

When comparing the baseline methods, we observed that in some cases, GS performs better than the original data. We notice that the 'Relevant' class is underrepresented in the data, which leads the model not to predict the 'Relevant' class. Therefore, this experiment aims to understand how each technique, GS, Extractive Summarization, Synonym Augmentation, and Semi-Automatic Labeling, can predict each class. Figure 2 shows the confusion matrix that illustrates and summarizes their classification performance. Looking at GS, we can see that the 'Relevant' class was misclassified as the 'Same Entities' class. In contrast, the other approaches can successfully predict the 'Relevant' class. We can also observe that Extractive Summarization performs the worst in predicting the 'Relevant' class. This is due to the distribution shift between original comments written by users and comments generated by article summarization. Finally, it is clear that all techniques can correctly identify irrelevant classes compared to the rest of the classes; this is an indication of clear separation between the 'Irrelevant' and the 'Relevant', 'Same Entities', and 'Same Categories' examples.

B. Q2: EFFECT OF HUMAN-IN-THE-LOOP AND UNLABELED DATA

1) EFFECT OF HUMAN CORRECTION

This experiment aims to explore the extent of the error rate λ in the Semi-Automatic Labeling technique and characterize the effect of λ on the model performance and human-in-the-loop time. First, we investigate the effect of λ on two extreme

TABLE 2. Test accuracy (in percent) for all techniques. The data are ordered based on the sequence length from smallest to largest. We report test accuracy and standard deviation calculated by repeating each experiment five times.

Source	Baselines						Proposed
	Original	GS	Traditional		Synthetic Generation		Semi-Automatic Labeling
			Oversampling	Weighted Loss	Extractive Summarization	Synonym Augmentation	
WSJ	63.1(2.2)	62.3(4.0)	58.1 (1.8)	64.6 (3.7)	68.1 (1.8)	75.4 (0.7)	**80.4 (0.1)**
FN	75.6(1.6)	82.5(1.1)	70.9 (5.8)	71.7 (4.0)	80.4 (3.0)	82.5 (1.5)	**92.2 (0.9)**
DM	67.3(3.1)	68.0(1.8)	64.6 (2.3)	64.3 (2.4)	67.4 (2.3)	71.4 (2.3)	**83.0 (0.2)**
TG	74.5(5.8)	74.5(2.2)	70.5 (1.5)	72.5 (2.6)	77.7 (2.6)	82.8 (1.9)	**84.5 (0.4)**
MW	75.2(4.1)	77.4(4.6)	68.4 (2.5)	71.8 (3.2)	72.4 (1.0)	80.5 (0.9)	**80.9 (0.1)**

cases when the annotators' agreement is the highest and lowest for our data, more specifically, when the agreement is highest for the FN source and lowest for the WSJ sources. The number of rounds in which humans need to manually correct data points appears to be correlated with λ; a smaller λ suggests that humans need a few additional rounds compared to using a higher λ. We observed that for all sources, a λ value between 9% and 7% required $1-2$ rounds of corrections, while a λ between 1% and 3% required at least 5 rounds of corrections. On the other hand, $\lambda = 5\%$ results in a practically reasonable number of correction rounds $(2-3)$.

Regarding the effect of λ on the model performance, in general, model performance and λ seem to have an inverse relationship (shown in Figure 4 (a)). In other words, a higher λ leads to lower performance and vice versa. For WSJ, we can see a peak in performance after decreasing λ from 9% to 7%; WSJ has the lowest agreement between annotators, and after two rounds of corrections, the model performance increases by 8%. However, on both WSJ and FN, the model performance becomes stable after $\lambda = 5\%$, which indicates that 5% is the most effective and efficient error rate that can yield high performance with reasonable cost since $\lambda = 5\%$ requires between $2-3$ rounds of human corrections compared to smaller λ. It is not surprising that more interaction between humans and the Semi-Automatic Labeling leads to higher costs, which we observed in Figure 4 (b) where $\lambda = 1\%$ requires much more human time (T_h) to make a proper correction, comparing to $\lambda = 9\%$.

2) EFFECT OF AMOUNT OF UNLABELED DATA

The purpose of this experiment is to explore to what extent the amount of unlabeled data affects the performance of the Semi-Automatic Labeling technique. We compared the accuracy in percentages for each source when we do not use any unlabeled data 0%, and once the number of unlabeled data is increased by 20%, 40%, 60%, 80%, and 100% of the total number of unlabeled data. The ratio between labeled and unlabeled data when the fraction of unlabeled data change is shown in Table 3. Figure 3 shows the test accuracy for different fractions of unlabeled data, where 0% indicates the model performance on GS data, without including the Semi-Automatic Labeling technique. We can see that the number of unlabeled examples plays a key role in gradually enhancing the performance obtained on all sources. For all sources, the

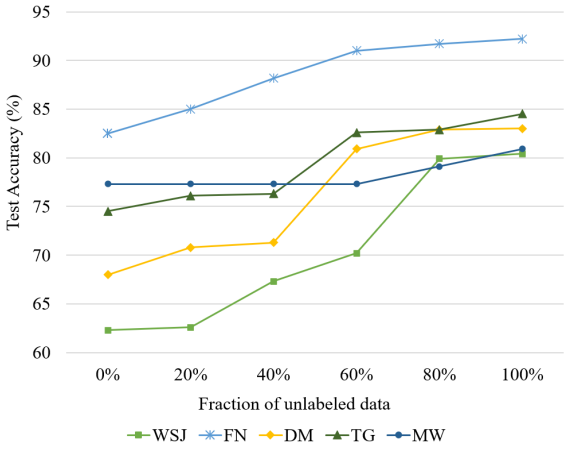


FIGURE 3. Test accuracy in percentages for each proportion of unlabeled data. 0% indicates that we only use GS. 100% means that we utilize all unlabeled data in a given source.

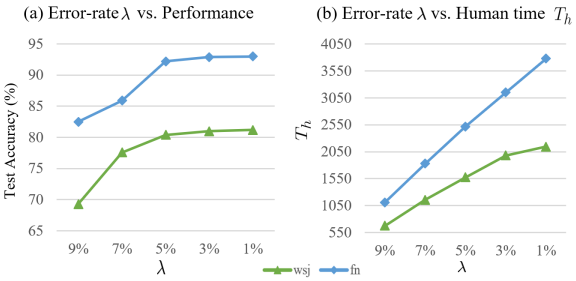


FIGURE 4. (a) The effect of error rate (λ) on the model performance. (b) The error rate (λ) and human correction time (T_h) in minutes for the Semi-Automatic Labeling approach on the FN and WSJ sources.

accuracy increases when unlabeled data is 4 times larger than labeled data. For WSJ and MW the peak occurs when the fraction of used unlabeled data is between $60\% - 80\%$ and $40\% - 60\%$ for DM, TG, and FN. The difference in peak starting point is due to the fact the total number of unlabeled examples in WSJ and MW, is much smaller than in other sources.

C. Q3: PERFORMANCE OF SYNTHETIC GENERATION TECHNIQUES

This experiment aims to understand the extent of synthetic generation on the model performance. We compare

TABLE 3. Ratio between labeled:unlabeled examples for each fraction. Note that 100% represents the usage of all unlabeled examples in a source, and the percentages 20% – 80% indicates the fraction of unlabeled examples that is used.

Source	Fraction of unlabeled examples (%)				
	20%	40%	60%	80%	100%
WSJ	1 : 1.40	1 : 2.70	1 : 4.00	1 : 5.50	1 : 6.90
FN	1 : 4.20	1 : 8.40	1 : 12.7	1 : 17.0	1 : 21.2
DM	1 : 7.20	1 : 14.4	1 : 21.6	1 : 28.8	1 : 36.0
TG	1 : 7.50	1 : 15.0	1 : 22.6	1 : 30.1	1 : 37.0
MW	1 : 1.02	1 : 1.90	1 : 3.00	1 : 4.00	1 : 5.10

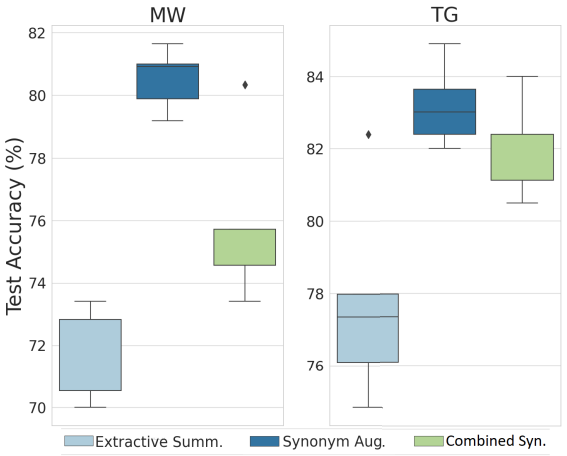


FIGURE 5. Boxplots of the test accuracies obtained by Extractive Summarization, Synonym Augmentation, and Combined Synthetic Generation for the MW and TG sources.

Extractive Summarization, where the generated comments are based on their respective articles; Synonym Augmentation, where the generated comments are based on original comments; and the Combined Synthetic techniques, where we combined both Extractive Summarization and Synonym Augmentation. As shown in Figure 5, Synonym Augmentation outperforms Extractive Summarization. This is due to the shift between the original and synthetic data (Synonym Augmentation produces synthetic data from the same distribution as the original comments). Moreover, Synonym Augmentation outperforms the Combined Synthetic technique, although the latter leverages more examples since it generates synthetic data using two techniques. We observed that Synonym Augmentation is more stable than the other two synthetic generation techniques. This is because Extractive Summarization produces some noise that confuses the model and makes it unstable when the combined technique is used as well.

D. Q4: MARGINAL BENEFIT VS. MARGINAL COST

The objective of Q4 is to aid decisions on balancing between performance and resource allocation. Although the Semi-Automatic Labeling approach obtains the highest performance across all sources, it is the most resource-consuming approach. This is due to two reasons. First, in this approach, the human factor t_h is considered in calculating the cost,

TABLE 4. Utility function values for each source. Based on Eq. (3), the larger the value, the better (highlighted in bold).

Source	Extractive Summ.	Synonym Aug.	Semi-Auto. Labeling
WSJ	0.63	**0.75**	0.43
FN	0.78	**0.82**	0.16
DM	0.64	**0.71**	0.37
TG	**0.77**	0.76	0.33
MW	0.70	**0.80**	0.42

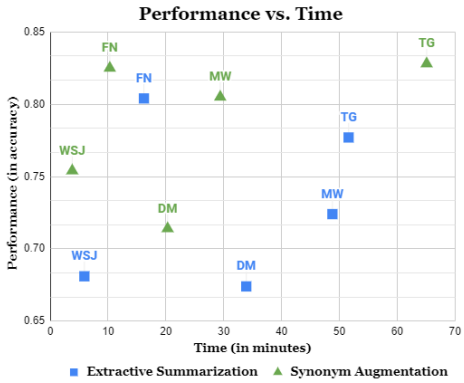


FIGURE 6. Performance vs. Time for Extractive Summarization and Synonym Augmentation.

while in Extractive Summarization and Synonym Augmentation, $t_h = 0$ since no human time is involved. Second, Semi-Automatic Labeling includes unlabeled data at least 7 times larger than the original data. Consequently, t_{cpu} is also larger in this case. Table 4 shows that when penalizing the approaches by the cost they consume using Eq. (3), Semi-Automatic Labeling yields the worst utilization leaving Extractive Summarization and Synonym Augmentation as better options from the perspective of utility. Further, looking closely, we can see that Synonym Augmentation manifests the best utilization except for TG, where the difference between the two approaches is insignificant.

Looking closely at Extractive Summarization and Synonym Augmentation, Figure 6 shows the time consumed (x-axis) against performance in terms of accuracy (y-axis) for both techniques. Each technique is depicted in different colors, while each point represents a different source. We can see that Synonym Augmentation yields better performance and less time compared to Extractive Summarization in all sources. Except for TG, where Synonym Augmentation is insignificant and time-consuming. This is because the average number of comment tokens in TG is 46, which means that Synonym Augmentation produces comments with larger sequences that consume more of the CPU time. On the other hand, Extractive Summarization produces comments that are between $15 - 20$ tokens which is 2 times shorter than comments generated by Synonym Augmentation.

In summary, with time and budget for labeling some article-comment pairs and access to larger computing resources, a practitioner should use Semi-Automatic

Labeling since it provides the best overall performance. However, in the case of no re-labeling budget and limited computing resources, a fully-automated Synonym Augmentation may be an alternative approach.

VII. CONCLUSION

This study proposed a framework for selecting an effective and efficient technique to address the imbalance in article-comment alignment due to annotators' disagreement given a constrained budget and desired performance. Our study shows that Semi-Automatic Labeling, where we utilize unlabeled data with a human-in-the-loop approach, outperforms all other techniques considered. It is a more expensive approach than fully automated alternatives that do not involve human data relabeling. When human-in-the-loop is not an option, we found that Synonym Augmentation, which generates new comments from the same distribution as the real comments, performs better than Extractive Summarization, which utilizes articles to generate new comments. We also conclude that Synonym Augmentation provides a more reasonable tradeoff between cost and performance with no time or budget for relabeling some article-comment pairs.

We observe that in the embedding space, the three classes 'Relevant', 'Same Entities', and 'Same Category' are closer to each other with some overlap. This phenomenon raises an important question of whether a multi-label problem where each article-comment pair receives multiple classes as a label is more suited for ACAP. On the other hand, we observed that the classes are more distinguishable in the case of some sources compared to others. Therefore, the question is if a systematic knowledge transfer between sources, where we learn from classes with the highest performance and then transfer this knowledge to other sources, will help the model to identify each class better. We keep the answer to these questions for future directions.

REFERENCES

[1] S. Akkaradamrongrat, P. Kachamas, and S. Sinthupinyo, "Text generation for imbalanced text classification," in *Proc. 16th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2019, pp. 181–186.

[2] J. Alshehri, M. Stanojevic, E. Dragut, and Z. Obradovic, "Stay on topic, please: Aligning user comments to the content of a news article," in *Proc. Adv. Inf. Retr.*, 2021, pp. 3–17.

[3] A. Anand, G. Pugalenthi, G. Fogel, and P. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," in *Proc. Amino Acids*, 2010, pp. 1385–1391.

[4] B. Beigman Klebanov and E. Beigman, "From annotator agreement to noise models," in *Proc. ACL*, 2009, pp. 495–503.

[5] E. Brangbour, P. Bruneau, T. Tamisier, and S. Marchand-Maillet, "Active learning with crowdsourcing for the cold start of imbalanced classifiers," in *Proc. Int. Conf. Cooperat. Design, Vis. Eng.*, 2020, pp. 192–201.

[6] F. Cheng, J. Zhang, and C. Wen, "Cost-sensitive large margin distribution machine for classification of imbalanced data," in *Pattern Recognit. Lett.*, vol. 180, pp. 107–112, Sep. 2016.

[7] A. Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," in *Proc. TACL*, 2022, pp. 92–110.

[8] M. Desmond, C. Finegan-Dollak, J. Boston, and M. Arnold, "Label noise in context," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2020, pp. 157–186.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 1–16.

[10] T. Dopierre, C. Gravier, J. Subercaze, and W. Logerais, "Few-shot pseudo-labeling for intent detection," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4993–5003.

[11] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[12] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[13] J. Ürnkranz and P. Flach, "An analysis of rule evaluation metrics," in *Proc. ICML*, 2003, pp. 202–209.

[14] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[16] L. Hou, J. Li, X. Li, J. Tang, and X. Guo, "Learning to align comments to news topics," *ACM Trans. Inf. Syst.*, vol. 36, no. 1, pp. 1–31, 2017.

[17] P. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: A study of annotation selection criteria," in *Proc. Workshop Active Learn. Natural Lang. Process. (NAACL HLT)*, 2009, pp. 1–9.

[18] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proc. NAACL*, 2018, pp. 1–6.

[19] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, pp. 221–232, Apr. 2016.

[20] D. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn. (ICML)*, 2013, pp. 1–6.

[21] G. Lemaitre, F. Nogueira, and C. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, pp. 1–5, Jan. 2017.

[22] Y. Li, G. Sun, and Y. Zhu, "Data imbalance problem in text classification," in *Proc. 3rd Int. Symp. Inf. Process.*, Oct. 2010, pp. 301–305.

[23] J. Madhuri and R. G. Kumar, "Extractive text summarization using sentence ranking," in *Proc. Int. Conf. Data Sci. Commun.*, 2019, pp. 1–3.

[24] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 1–27, Nov. 2008.

[25] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining Knowl. Discovery*, vol. 28, no. 1, pp. 92–122, Jan. 2014.

[26] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *Proc. Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, Jan. 2017, pp. 1–6.

[27] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021.

[28] R. Németh, D. Sik, and F. Máté, "Machine learning of concepts hard even for humans: The case of online depression forums," *Int. J. Qualitative Methods*, vol. 19, pp. 1–8, Jan. 2020.

[29] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," *Proc. Comput. Sci.*, vol. 159, pp. 736–745, Jan. 2019.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," in *Proc. Stanford InfoLab*, 1999, pp. 1–5.

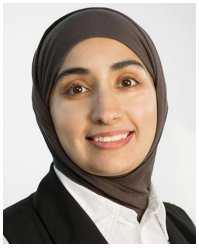
[31] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–12.

[32] D. K. Sil, S. H. Sengamedu, and C. Bhattacharyya, "ReadAlong: Reading articles and comments together," in *Proc. 20th Int. Conf. Companion World Wide Web*, Mar. 2011, pp. 125–126.

[33] A. Sun, E. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decis. Support Syst.*, vol. 48, no. 1, pp. 191–201, Dec. 2009.

[34] A. Tharwat and W. Schenck, "A novel low-query-budget active learner with pseudo-labels for imbalanced data," in *Mathematics*, vol. 10, no. 7, p. 1068, Mar. 2022.

[35] J. Zhang, X. Wu, and V. S. Shengs, "Active learning with imbalanced multiple noisy labeling," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1095–1107, May 2015.



JUMANAH ALSHEHRI received the B.Sc. degree in management information systems, the master's degree in business administration from King Faisal University, KSA, and the M.Sc. degree in computer science from Saint Joseph's University, USA. She is currently pursuing the Ph.D. degree in computer and information sciences with Temple University, USA, under the supervision of Dr. Zoran Obradovic. Her main research interests include machine learning, natural language processing, text mining, and social and information networks. Her research revolves around modeling and extracting knowledge from textual data, specifically, jointly mining and modeling articles and user-generated content and then utilizing it in downstream applications such as article-comment semantic analysis, article-comment alignment problem, media bias classification, and fake news detection.

MARTIN PAVLOVSKI received the B.Sc. degree in electrical engineering and information technologies from the Saints Cyril and Methodius University of Skopje, Skopje, Macedonia, in 2015, and the Ph.D. degree in computer and information science from Temple University, Philadelphia, PA, USA, in 2021. He was a Researcher with the Macedonian Academy of Sciences and Arts and a Visiting Scholar at Temple University. He is currently a Research Scientist with Yahoo. His research interest includes machine learning from structured data.

EDUARD DRAGUT (Member, IEEE) received the Ph.D. degree in computer science from the University of Illinois, Chicago, in 2010. He is currently an Associate Professor with the Computer and Information Sciences Department, Temple University. He is the coauthor of the book *Deep Web Query Interface Understanding and Integration*. His research interests include web-based information retrieval, cleaning, integration, and mining. He has co-chaired multiple workshops, including the series on "Data Science with Human in the Loop (DaSH)."

ZORAN OBRADOVIC (Senior Member, IEEE) is currently a Distinguished Professor, the Center Director of Temple University, an Academician with the Academia Europaea (the Academy of Europe), and a Foreign Academician with the Serbian Academy of Sciences and Arts. He has mentored 45 postdoctoral fellows and Ph.D. students, many of them have independent research careers at academic institutions and industrial research laboratories. His research results were published in about 400 data science and complex networks articles addressing challenges related to big, heterogeneous, spatial–temporal data analytics motivated by applications in healthcare management, power systems, earth, and social sciences. He is an editorial board member of 13 journals. He was a general chair, program chair, or track chair of 11 international conferences. He is the Steering Committee Chair of the SIAM Data Mining Conference. He is the Editor-in-Chief of the *Journal of Big Data*.