On Label Quality in Class Imbalance Setting - A Case Study

Jumanah Alshehri*
Temple University
Philadelphia, PA, USA
shehri.j@temple.edu

Marija Stanojevic*
Temple University
Philadelphia, PA, USA
marija.stanojevic@temple.edu

Eduard Dragut*
Temple University
Philadelphia, PA, USA
edragut@temple.edu

Zoran Obradovic*
Temple University
Philadelphia, PA, USA
zoran.obradovic@temple.edu

Abstract*—Producing high-quality labeled data is a challenge in any supervised learning problem, where in many cases, human involvement is necessary to ensure the label quality. However, human annotations are not flawless, especially in the case of a challenging problem. In nontrivial problems, the high disagreement among annotators results in noisy labels, which affect the performance of any machine learning model. In this work, we consider three noise reduction strategies to improve the label quality in the Article-Comment Alignment Problem, where the main task is to classify article-comment pairs according to their relevancy level. The first considered labeling disagreement reduction strategy utilizes annotators' background knowledge during the label aggregation step. The second strategy utilizes user disagreement during the training process. In the third and final strategy, we ask annotators to perform corrections and relabel the examples with noisy labels. We deploy these strategies and compare them to a resampling strategy for addressing the class imbalance, another common supervised learning challenge. These alternatives were evaluated on ACAP, a multiclass text pairs classification problem with highly imbalanced data, where one of the classes represents at most 15% of the dataset's entire population. Our results provide evidence that considered strategies can reduce disagreement between annotators. However, data quality improvement is insufficient to enhance classification accuracy in the article-comment alignment problem, which exhibits a high-class imbalance. The model performance is enhanced for the same problem by addressing the imbalance issue with a weight loss-based class distribution resampling. We show that allowing the model to pay more attention to the minority class during the training process with the presence of noisy examples improves the test accuracy by 3%.

***Index Terms*—Label quality, annotators disagreement, data imbalance problem, multiclass text classification**

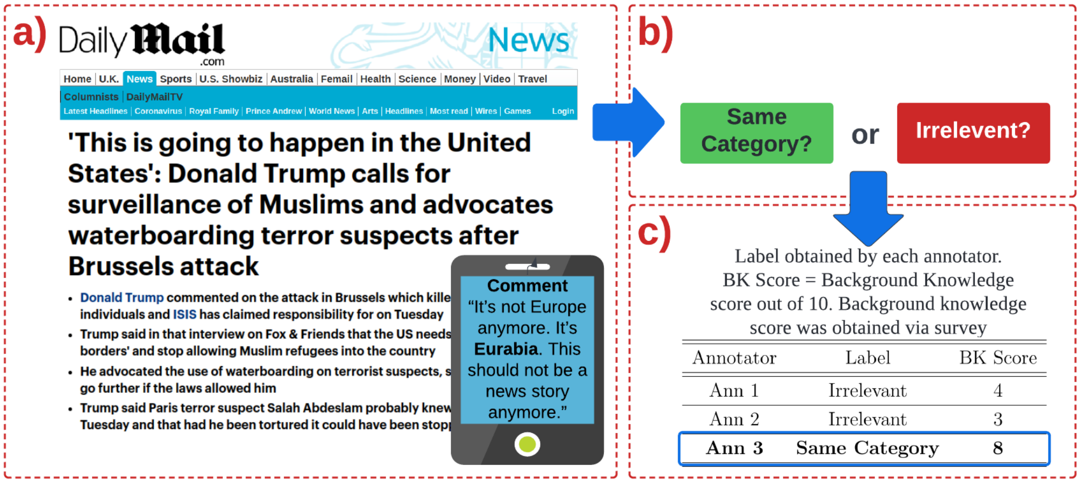
I. Introduction

Human labels (annotations) produces high quality labels , however, the process can vary from straightforward to complex, depending on the problem. For instance, categorizing the relevancy between a news article and their associated comments requires reading comprehension and familiarity with the background of the story in the article. Label quality is not the only challenge in in supervised learning; sampling is also a major challenge (e.g., class imbalance). In the case of low resources problems, the quality of labeled data is crucial. Obtaining high-quality labels through experts or human annotations is costly and such annotations are not error-free. To address the issue of label quality, one approach is to elicit multiple annotators to label the data, then take the mean or majority vote of the obtained labels. There are tasks where multiple annotators may produce different labels leading to high disagreement. This introduces the problem of noisy labels [5], [7], [21], which adversely affects the ability of training robust models. Hiring expert annotators is one way to gather high quality data; however, expert annotation is expensive. An alternative is to train annotators by giving them clear and unambiguous instructions avoiding introducing bias. One also need to consider the role annotators' background knowledge may play in the labeling process. For instance, labeling articles and comments that discuss foreign political issues might not be easy for someone who does not follow foreign affairs. In the example shown in Figure 1, we asked three annotators to label the article-comment pairs[1] as "Same Category" or "Irrelevant". In this example two annotators rate the comment as "Irrelevant", while the third annotator rates it as "Same Category". The appropriate label is "Same Category", but one requires familiarity (background knowledge) with the political circumstances in the Arab and European countries in 2016 to assign that label. Taking the majority vote, the aggregated label becomes "Irrelevant," which is incorrect, illustrating that we need to consider annotators' background knowledge to obtain the correct final label.

In this work, we discuss our encounter with those issues in a concrete application, aligning user comments to the content of a news article. It has been shown that irrelevant comment discovery is helpful in user comment summarization [12] and also in detecting topic drifts [13]. We report several strategies to enhance the predictions in the Article-Comment Alignment Problem (ACAP) [1] in the presence of high user annotation disagreement and class imbalance. We discuss ACAP in detail in Section III. In our setting, we encounter two main challenges: 1) noisy label, caused by the high disagreement among annotators since, as we discovered firsthand, users often have difficulty stating if a comment is related to the story presented in an article. 2) sampling user comments to be labeled by human annotators which gives highly imbalanced datasets.

Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA.

[1] Full article: https://dailym.ai/2Qz7RG9


a)
Daily Mail
.com
News
Home | U.K. | News | Sports | U.S. Showbiz | Australia | Femail | Health | Science | Money | Video | Travel
Columnists | DailyMailTV
Latest Headlines | Coronavirus | Royal Family | Prince Andrew | World News | Arts | Headlines | Most read | Wires | Games
Login
b)
Same Category?
or
Irrelevent?
c)

'This is going to happen in the United States': Donald Trump calls for surveillance of Muslims and advocates waterboarding terror suspects after Brussels attack

- Donald Trump commented on the attack in Brussels which kille individuals and ISIS has claimed responsibility for on Tuesday
- Trump said in that interview on Fox & Friends that the US needs borders' and stop allowing Muslim refugees into the country
- He advocated the use of waterboarding on terrorist suspects, s go further if the laws allowed him
- Trump said Paris terror suspect Salah Abdeslam probably knew Tuesday and that had he been tortured it could have been stopp

Comment
“It’s not Europe anymore. It’s **Eurabia**. This should not be a news story anymore.”

Label obtained by each annotator.
BK Score = Background Knowledge score out of 10. Background knowledge score was obtained via survey

Annotator	Label	BK Score
Ann 1	Irrelevant	4
Ann 2	Irrelevant	3
Ann 3	**Same Category**	**8**

Fig. 1. Article and comment pair from Daily Mail. a) Part of the article is shown and the desired comment to be labeled. b) The labels the annotators have to choose based on their understanding of the relationship between the article-comment pairs. c) Labels obtained by annotators and the Background Knowledge Score (BK Score) represent their confidence level regarding the article topic. The most accurate label is the label obtained by the third annotator (Ann 3).

II. Related Work

Learning from low quality labels is a common challenge in machine learning practice. Several works have used different strategies to overcome this problem. Some research focus on diminishing the effect of noisy labels by enlarging the number of data points through use of weak labels and crowdsourcing [18]–[20], [28]. Another line of work utilizes active learning, where the algorithm samples the desired data to be labeled by the annotators [22], [23]. Some research focuses on improving label quality through a noise correction mechanism that model and correct the noise labels using the estimated quality of the ground truth labels [9], [24], [25].

This work aims to find an effective strategy that utilizes annotators' disagreement and background knowledge to improve the label quality and compare that with relabeling for ACAP. In challenging applications [15], [17], [26], it is often difficult to get high-quality labels. Our work is different since we are not only facing high inner disagreement among annotators but also high-class imbalance.

III. Problem Formulation

A. Article-Comment Alignment Problem

ACAP's primary goal is to classify an article-comment pair based on the relevancy level. We classify each article-comment pair into one of four classes. 1) *Relevant* - the content of the comment discusses the same matter as the article; for example, the article discusses various issues that affected the world due to the Russian-Ukraine War, such as the flow of goods, fueling dramatic cost increases and product shortages. The comment that discusses how the current increase in oil prices due to the war affected them is relevant. 2) *Same Category*- the comment does not discuss the same issue as the article, but it is within the same scope; for instance, the article discusses the 2022 Russian invasion of Ukraine, and the comment describes some incidence from the Slovenian War in 1991. The article and comment do not discuss the same topic but are within the same scope. 3) *Same Entities*- the article-comment pair are from this class if they are Irrelevant, but the same entities are mentioned. For example, the article discusses the United States' perspective regarding the Russian-Ukraine War, mentioning President Joe Biden. At the same time, the comment focuses on president Joe Biden's speech in Pittsburgh-area union hall during labor day. Here, the article and comment mention the same entity in different contexts. 4) *Irrelevant*- the comment content is irrelevant to the article and does not belong to any of the above classes. In addition to finding the relevancy level for an article-comment pair, ACAP helps in understating that type of relevancy and uses it as a filtering tool for other downstream applications [2]. For example, ACAP may help analyzing commenters' opinions regarding entities discussed in the article and how opinion drifts over time.

B. Data and Labeling

The articles and their comments were collected between year 2015 and 2017 [10] from five news outlets - Wall Street Journals (WSJ), Fox News (FN), Daily Mail (DM), The Guardian (TG), and Market Watch (MW), where each outlet represents a dataset. Each dataset consists of around 1K labeled article-comment pairs. We involve three English speaking undergraduate students as annotators to manually label each article-comment pair according to the pre-defined classes. We provide annotators with the following information: 1) article-comment pairs without the surrounding context, and 2) specific description for the four proposed classes, with no specific order or degree of intensity to avoid bias. Each class was associated with score {0, 1, 2, 3}, corresponding to {Irrelevant, Same Entity, Same Category, and Relevant}, respectively. We trained annotators individually by explaining each class's definition and providing some examples that were not part of the final dataset. To avoid bias, the annotators were unaware of each other, and the labeling process was independent. Then, we aggregated the scores given by each annotator for each article-comment pair using an averaging aggregation schema. The initially labeled data is highly imbalanced in one class, the "Relevant", where it represents a distribution from 3%-15% in specific datasets. The remaining classes distribution varies between 20%-51% of the entire population, making the dataset highly imbalanced in one class compared to the remaining three classes.

C. Agreement Analysis

To measure the inner agreement among annotators, we utilize two annotators agreement metrics: 1) Fleiss Kappa statistic [8], 2) Krippendorff's alpha coefficient [11]. *Fleiss Kappa statistic* calculates agreement between multiple scorers on each class and then averages the scores over classes to produce the final statistics.

To account for the magnitude of error a scorer makes, we use *Krippendorff's alpha coefficient* to calculate overall agreement in labeling. This statistic works by considering the distance between labels given by multiple scorers. It is calculated by subtracting the disagreement among assigned values normalized by the disagreement achieved if labels are assigned by chance. $\alpha \in [0, 1]$, where 0 indicates random scoring, and 1 indicates a perfect correspondence between annotators.

Table I shows the agreement scores for the three annotators for each datasets. Results show that labeling article-comment pairs in WSJ are the most challenging task, with the smallest correspondence between the annotators. The raters' agreement is "Fair" for WSJ, TG, DM, and MW and borderline "Moderate" for FN, according to the Fleiss Kappa statistic. The Krippendorff's alpha score[2] is between 42% and 66% across the outlets. Both metrics indicate the difficulty humans have when assigning the class of comment in general.

TABLE I
AGREEMENT ANALYSIS FOR LABELING ARTICLE-COMMENT PAIRS IN FIVE NEWS OUTLETS. FK= FLEISS KAPPA STATISTIC, α= KRIPPENDORFF'S ALPHA COEFFICIENT.

Dataset	**WSJ**	**TG**	**DM**	**MW**	**FN**
FK	0.22	0.36	0.37	0.40	0.45
α	0.42	0.60	0.61	0.64	0.66

D. Gold and Noisy Labels

The high disagreement among annotators produces noisy labels, which leads to poor performance of the model. To solve this problem, we flag each example in our datasets to Gold Label (GL) and Noise Label (NL). The flagging criteria are based on the inter-annotator variation σ. A $\sigma = 0$ indicates that all annotators agree on one label, and this example is flagged as GL. On the other hand, $\sigma > 0$ means that at least one annotator disagrees with the other annotators, and this example will consider NL. Table II show the proportion of GL and NL examples in each dataset. Note that the number of GL and NL varies across the datasets, where WSJ has the highest percentage of NL 56%, compared to the other datasets, where the NL percentage is only between $14\% - 20\%$. This behavior will help better investigate the effect of NL on the model performance.

TABLE II
THE RATIO OF GL (GOLD LABEL) AND NL (NOISY LABEL) EXAMPLES IN EACH DATASET

Dataset	**# GL**	**% GL**	**# NL**	**% NL**
WSJ	443	44%	557	56%
TG	795	80%	205	20%
DM	833	83%	167	17%
FN	858	86%	142	14%
MW	862	86%	138	14%

IV. METHODOLOGY

In this work, we consider three different strategies to reduce the problem of noisy labels, where all strategies revolve around the annotators. We revert to the annotators by utilizing the annotators' background knowledge and agreement scores and relabeling some NL examples.

A. Pre-Weight Labeling (Pre-WL)

Annotators' knowledge differs based on their interests. For example, an annotator might be interested in politics but not entertainment, making her more knowledgeable in political news than news related to movies and actors. Therefore, this technique uses the annotators' background knowledge to obtain the final label. The paper review process in journals and conferences inspires this strategy, because reviewers are asked to provide their confidence level (e.g., Expert, Knowledgeable, and Familiar). Therefore, we asked annotators to scale their knowledge regarding each news topic (international politics, national politics, international sports, national sports, international entertainment, national entertainment, health, business, science, and technology) from [1-10]; 1 being not knowledgeable and 10 being very knowledgeable. Next, we weight each label given by an annotator with her confidence level based on her knowledge scale (λ). We calculate confidence level (ω) as follows:

$$\omega = \begin{cases} 0.3 & \text{if } \lambda = [1-4] \\ 0.6 & \text{if } \lambda = [5-7] \\ 1 & \text{if } \lambda = [8-10] \end{cases}$$

The interpretation of confidence level is as follow, $\omega = 0.3$ represents low confidence, $\omega = 0.6$ represents medium confidence, and $\omega = 1$ represents high confidence. In other words, for each article-comment pair and in the label aggregation process, the label obtained by an annotator with high confidence will be weighted more than the label obtained by a low confidence annotator. The final aggregated label is calculated as shown at Equation 1:

$$\hat{y} = \sum_{i=1}^{n} (l_i \, \omega_{ic})/n \qquad (1)$$

[2] Calculted by http://dfreelon.org/utils/recalfront/recal-oir/ software using ordinal setting

Here, $\hat{y}$ is the aggregated label, n is the number of annotators, l_i is the label given by the i^{th} annotator, and ω_{ic} is the confidence level for the i^{th} annotator for a given topics c. Using this equation, we recalculate the aggregated label for all NL examples. Once the aggregated labels are obtained, we train and test the model with GL and NL with the new Pre-WL labels.

B. Post-Weight Labeling (Post-WL)

This strategy utilizes the inner-disagreement between annotators by allowing the model to treat each example differently during the training process according to the disagreement level. In other words, examples with low disagreement will contribute more to the learning process than examples with high disagreement. To achieve this goal, we first calculate σ, which is the inner-annotator variation per example. Next, calculate the corresponding weight (ω) for each example by leveraging the exponential growth and decay concept; (ω) will be calculated as follow, (ω) = 1 if σ = 0, and δ^{σ}, otherwise.

Here, δ is a hyper-parameter representing the change rate. We then integrate ω with the model loss function. Thus, examples with low disagreement will obtain an $\omega = 1$, forcing the model to be more attentive to that examples. Contrary, examples with high disagreement will receive a lower ω which will force the model to pay less attention to that example. The true value of ω is $0 < \omega \leq 1$. We tried different values of δ and found via experiments that 0.5 give the best performance.

C. Annotator Relabeling

This approach focuses on reverting to the annotators to relabel the NL examples. We start by identifying NL examples and understanding annotators' common mistakes during labeling. Then, we meet with the annotators, explain the labeling mistakes using a few NL examples, and ask them to relabel the examples without looking into the previous noisy label. Once the NL examples are relabeled, we recalculate the users' agreement score, train the model on GL, and relabel NL examples. Figure 2 shows the agreement score using Fleiss Kappa statistics and Krippendorff's α coefficient before and after relabeling NL. As shown, the inner-agreement score increased in both metrics in all datasets, especially in WSJ, where the agreement score was the lowest before relabeling.

V. Experimental Setup

A. Baselines:

1) Original: In the baseline strategy, we train and test on GL and NL without making any changes to NL.

2) Gold Labels (GL): Here, we exclude the NL examples and only train and test on GL examples. This strategy might seem ideal in the case of many high quality labeled examples. However, the question is, how will this strategy perform with a small number of training examples (less than $1K$)?

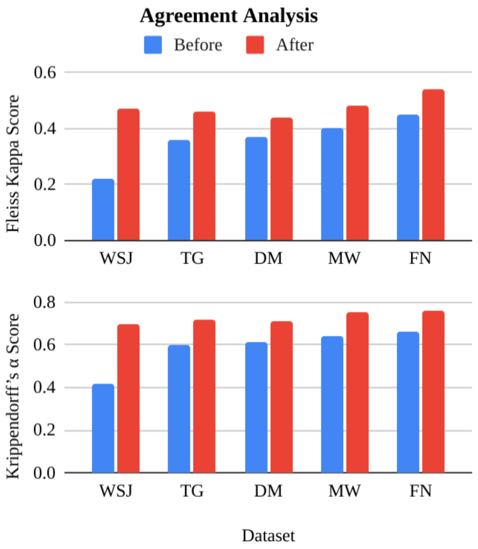


Fig. 2. Agreement score before and after relabeling NL, using two metrics Fleiss Kappa statistics and Krippendorff's α.

3) Random Labeling (RL): In this naive strategy, we randomly assign a label for the NL examples that are different from the given noisy label. For example, if the noisy label is 1, we randomly assign 2, 3, or 4 to that example. Then we train and test the model on union of GL and randomly relabeled NL. The problem with this approach is that if we randomly select the most appropriate label, there is little logical explanation to support the findings.

B. Model:

We apply BERTAC [1], a BERT [6] based model that jointly learns articles and comments in an end-to-end fashion. BERTAC allows the model to learn more expressive embeddings that encode the relevance between an article and its comment. We repeat experiments five times on different randomized splits for all baselines and other strategies. The dataset split is a 70:20:10 ratio for training, testing, and cross-validation. We report the mean and standard deviation for all experiments.

Given our labels' ordinal nature, we introduce the ordinal classification loss, which accounts for the distance between the predicted class and the actual class. In the ordinal classification loss, we multiply BERTAC's loss for each example with a weight that is calculated according to Equation 2, where $k = 4$ (number of classes), y_i is the actual label, and $\bar{y}_i$ is the predicted label of the example.

$$weight = 1 + \frac{|\bar{y}_i - y_i|}{k - 1} \tag{2}$$

If the model predicted the correct class, the weight is 1; however, if the model predicts the wrong class, BERTAC classification loss is multiplied by 2, 3, or 4 based on the distance between the actual class and the predicted class.

TABLE III
TEST ACCURACY (IN PERCENT) AND STANDARD DEVIATION BETWEEN ROUND BRACKETS FOR ALL STRATEGIES. THE DATA ARE ORDERED BASED ON THE SEQUENCE LENGTH FROM SMALLEST TO LARGEST. BOLD VALUES REPRESENT THE HIGHEST TEST ACCURACY FOR EACH DATASET.

	Dataset	WSJ	FN	DM	TG	MW
Baselines	Original	**86.8(.4)**	91.5(.6)	88.6(.9)	90.5(.9)	91.3(.8)
	GL	85.0(.2)	**92.9(.3)**	**89.5(.5)**	**91.2(.8)**	**92.0(.9)**
	RL	64.7(.9)	71.8(.9)	66.4(.9)	63.8(.6)	72.5(.5)
Strategies	Pre-WL	84.3(.2)	81.7(.7)	82.7(.4)	66.7(.4)	86.2(.6)
	Post-WL	80.7(.4)	83.3(.6)	74.4(.9)	73.9(.4)	84.8(.7)
	Relabel	83.1(.9)	88.0(.2)	85.0(.3)	88.7(.6)	88.4(.1)

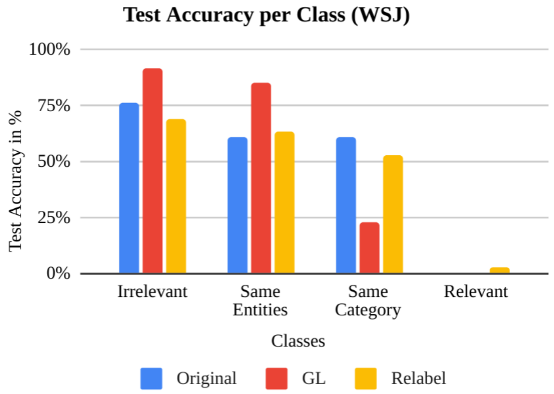


Fig. 3. Test accuracy in percentages per class for Original, GL, and Relabel strategy for the WSJ dataset.

VI. RESULTS AND DISCUSSION

The objective of conducted experiments was to determine if the data labeling quality improvement can overcome the data imbalance problem better as compared to an automated strategy relying on data resampling.

Table III shows the overall performance for baselines and compared labeling disagreement reduction strategies. In the first three rows, baselines, we can see that Random Labels (RL) perform the worst compared to Original and Gold Labels (GL). RL is a "pure luck" strategy with a probability of 33% that a particular random label matches the relevancy level between article-comment pairs. We can see that the GL outperforms the Original, which means that the noisy labels in the Original confuse the model; this affects the performance. However, this is not the case for WSJ, where removing the noise and only using GL examples dose not improve the performance; this is because NL examples in WSJ represent more than 50% of the dataset population compared to the rest of the dataset, and NL represents between 14%-20%. We hypothesize that this is because there are not enough examples for the model to learn from when using GL.

Surprisingly, none of those strategies, including relabeling, outperform Original and GL. Although the agreement score between annotators shown in Figure 2 increases between 9%-25% when relabeling the NL examples, the model performance in Relabel strategy declines between 2%-3% compared to the Original. This observation raises a new question: Is it reasonable to waste resources and relabel more examples? We analyze the performance in predicting each label to understand the model behavior. Figure 3 shows an example from WSJ for the model test accuracy performance for Original, GL, and Relabel. Although the overall test accuracy for Original and GL is preferable, with test accuracy of 86% and 85% respectively, we can see that both strategies fail in predicting the "Relevant" class. It can be explained by studying Table IV, the performance per class is proportional to the class distribution; for example, in the Relabeling strategy, the "Relevant" class distribution increased by 5% which allows the model to make correct predictions for that class. Also, for the "Same Category" class, the class distribution decreases in the GL, which leads to poor predictions. However, given the distinct semantics of each class, the "Same Entities" class performance is not affected much by the class distribution. The entity name in the article-comment pair helps the model learn this class better, even in the presence of few examples.

TABLE IV
CLASSES DISTRIBUTION IN PERCENTAGES FOR ORIGINAL, GL, AND RELABEL STRATEGY FOR THE WSJ DATASET.

Strategy	Original	GL	Relabel
Irrelevant	33%	54%	37%
Same Entities	34%	18%	24%
Same Category	27%	19%	25%
Relevant	6%	9%	14%

These observations change the problem directions; going back to the annotators and ask them to relabel the data is a misuse of resources in our case. In addition, this confirms that high-quality data alone will not overcome the sampling imbalance problem. We conduct a simple experiment to prove our assumption by choosing one of the traditional data imbalance methods, called Weighted Loss [4], using Original data. In this method, we modify the model loss function to account for the minority class by penalizing the model when predicting data points from the minority class during the training process. As shown in Table V, reducing the class imbalance problem with the Weighted Loss (W-Loss) method, while keeping noise labels, enhances the model performance.

TABLE V
TEST ACCURACY RESULTS IN PERCENTAGES BETWEEN ORIGINAL, GOLD LABELS (GL), RELABEL, AND WEIGHTED LOSS (W-LOSS) METHOD APPLIED ON ORIGINAL DATA.

Dataset	Original	GL	Relabel	W-Loss
WSJ	86.8	85.0	83.1	**88.6**
FN	91.5	92.9	88.0	**93.0**
DM	88.6	89.5	85.0	**89.7**
TG	90.5	91.2	88.7	**91.9**
MW	91.3	92.0	88.4	**92.5**

VII. Conclusion

In this work, we analyze several strategies for enhancing human annotators' label quality for the Article-Comment Alignment Problem (ACAP), where the goal is to understand the relevancy level between an article and it user comments. We consider three strategies to improve the quality of labeled data due to high disagreement between annotators. Two of the strategies require reverting to the annotators to 1) obtain more information regarding their background knowledge and 2) retrain annotators to relabel some of the examples. In the third strategy, we utilize the disagreement among annotators by incorporating the disagreement in the model loss function. The model gives higher weight to examples with high agreement scores and lower to examples with low agreement scores. Our results show that despite reducing the disagreement between annotators, in the case of imbalanced data, this does not help enhance the model's performance. Nevertheless, one needs to be aware of potential misuse of resources (like time and money). In contrast, we advocate that one needs to consider reducing class imbalance, in addition to allocating resources to relabeling, as this also can help enhance a model's overall performance. In the future, we will focus on combining data imbalance methods with our label quality strategies to further enhance the predictions of ACAP. We also plan to identify more problems with high class imbalance and noisy labels, and work through the lessons learned in this case study.

Acknowledgements

This work is supported in part by the U.S. NSF awards 2026513 and 1838145, and ARL subaward 555080-78055 under Prime Contract No.W911NF222000, 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 and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Alshehri, J., Stanojevic, M., Dragut, E. & Obradovic, Z. Stay on Topic, Please: Aligning User Comments to the Content of a News Article. *Advances In Information Retrieval*. (2021)

[2] Alshehri, J., Stanojevic, M., Khan, P., Rapp, B., Dragut, E. & Obradovic, Z. MultiLayerET: A Unified Representation of Entities and Topics using Multilayer Graphs. *Proceedings Of The ECML/PKDD*. (2022)

[3] Beigman Klebanov, B. & Beigman, E. From Annotator Agreement to Noise Models. *Association for Computational Linguistics*.(2009)

[4] Cao, K., Wei, C., Gaidon, A., Arechiga, N. & Ma, T. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. *Proceedings Of The 33rd International Conference On Neural Information Processing Systems*. (2019)

[5] Desmond, M., Finegan-Dollak, C., Boston, J. & Arnold, M. Label Noise in Context. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*.(2020)

[6] Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.(2019)

[7] Frenay, B. & Verleysen, M. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions On Neural Networks And Learning Systems*.(2014)

[8] Fleiss, J. Measuring nominal scale agreement among many raters.. *Psychological Bulletin*.(1971)

[9] Garg, S., Ramakrishnan, G. & Thumbe, V. Towards Robustness to Label Noise in Text Classification via Noise Modeling. *Proceedings Of The 30th ACM International Conference On Information & Knowledge Management*.2021)

[10] He, L., Han, C., Mukherjee, A., Obradovic, Z. & Dragut, E. On the Dynamics of User Engagement in News Comment Media. *WIREs*. (2019)

[11] Krippendorff, K. Computing Krippendorff's alpha-reliability. *Scholarly Commons*. (2011)

[12] Lu, Y., Zhai, C. & Sundaresan, N. Rated Aspect Summarization of Short Comments. *Proceedings Of The 18th International Conference On World Wide Web*. (2009)

[13] Mullick, A., Bhandari, A., Niranjan, A., Sckhar, N., Garg, S., Bubna, R. & Roy, M. Drift in Online Social Media. *2018 IEEE 9th Annual Information Technology, Electronics And Mobile Communication Conference*.(2018)

[14] Nettleton, D., Orriols-Puig, A. & Fornells, A. A Study of the Effect of Different Types of Noise on the Precision of Supervised Learning Techniques. *Artif. Intell. Rev.*(2010)

[15] Németh, R., Sik, D. & Máté, F. Machine Learning of Concepts Hard Even for Humans: The Case of Online Depression Forums. *International Journal Of Qualitative Methods*.(2020)

[16] Pechenizkiy, M., Tsymbal, A., Puuronen, S. & Pechenizkiy, O. Class Noise and Supervised Learning in Medical Domains: The Effect of Feature Extraction. *19th IEEE Symposium On Computer-Based Medical Systems*.(2006)

[17] Plank, B., Hovy, D. & Søgaard, A. Learning part-of-speech taggers with inter-annotator agreement loss. *Proceedings Of The 14th Conference Of The European Chapter Of The Association For Computational Linguistics*.(2014)

[18] Sheng, V., Provost, F. & Ipeirotis, P. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. (2008)

[19] Stanojevic, M., Alshehri, J. & Obradovic, Z. Surveying public opinion using label prediction on social media data. *ASONAM*. (2019)

[20] Stanojevic, M., Alshehri, J., Dragut, E. & Obradovic, Z. Biased news data influence on classifying social media posts. *NewsIR@ SIGIR*. (2019)

[21] Xiao, T., Xia, T., Yang, Y., Huang, C. & Wang, X. Learning from massive noisy labeled data for image classification. *2015 IEEE Conference On Computer Vision And Pattern Recognition*.(2015)

[22] Yan, Y., Huang, S., Chen, S., Liao, M. & Xu, J. Active learning with query generation for cost-effective text classification. *Proceedings Of The AAAI Conference On Artificial Intelligence*. (2020)

[23] Zhu, J., Wang, H., Yao, T. & Tsou, B. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. *Proceedings Of The 22nd International Conference On Computational Linguistics*. (2008)

[24] Zhang, J., Sheng, V., Wu, J., Fu, X. & Wu, X. Improving Label Quality in Crowdsourcing Using Noise Correction. *Proceedings Of The 24th ACM International On Conference On Information And Knowledge Management*. (2015)

[25] Zhang, J., Sheng, V., Li, T. & Wu, X. Improving Crowdsourced Label Quality Using Noise Correction. *IEEE Transactions On Neural Networks And Learning Systems*. (2018)

[26] Zhang, S., He, L., Dragut, L., Vucetic, S. How to invest my time: Lessons from human-in-the-loop entity extraction. *SIGKDD*. (2019)

[27] Zhu, X. & Wu, X. Class Noise vs. Attribute Noise: A Quantitative Study.. *Artif. Intell. Rev.* (2004)

[28] Zhang, J., Wu, X. & Shengs, V. Active Learning With Imbalanced Multiple Noisy Labeling. *IEEE Transactions On Cybernetics*. (2015)