From Tweets to Reddit: Leveraging Semi-supervised Domain Adaptation for Improving Data Filtering

Shelly Gupta[1](✉), Jumanah Alshehri[1,2], Ameen Abdel Hai[1], Hussain Otudi[1], and Zoran Obradovic[1]

[1] Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA
{shelly.gupta,shehri.j,aabdelhai,hussain.otudi,zoran.obradovic}@temple.edu
[2] Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia

Abstract. Reddit has emerged as a leading platform for microblogging data collection, providing valuable insights into patterns and knowledge discovery. However, the process of gathering and preparing this data presents significant challenges, particularly when it comes to ensuring its accuracy. Existing methodologies often yield an abundance of irrelevant posts, and datasets for relevance prediction on Reddit are rare. To overcome these obstacles, we propose a new semi-supervised framework that filters Reddit posts based on their topic relevance. Our approach combines annotated data from Twitter with weak labels generated from Wikipedia pages associated with relevant subreddits to automatically label Reddit posts. To enhance the model's generalization performance, we utilize a domain adversarial adaptation network to bridge the distribution gap between Twitter and Reddit data. Our novel framework achieves an accuracy of 73% and an F1 score of 0.77, which is a significant improvement of 20% compared to baseline models. Additionally, we address important research questions regarding the effectiveness of automatic labeling, the use of weakly labeled data, the contextual requirements for training domain adaptation models, and the optimal weak labeling method.

Keywords: Domain Adaptation · Reddit · Twitter · Semi-supervised · Relevance learning

1 Introduction

The internet has revolutionized global communication, with social media platforms emerging as pivotal sources of data over the past two decades. Microblogging platforms like Twitter, Facebook, and Reddit have become leaders in this area, with their usefulness in pattern recognition, knowledge discovery, information integration, scalability, and visualization being extensively studied [4]. While Twitter and Facebook are primarily used for short, real-time updates on

Published by Springer Nature Switzerland AG 2024
I. Maglogiannis et al. (Eds.): AIAI 2024, IFIP AICT 714, pp. 290–304, 2024.
https://doi.org/10.1007/978-3-031-63223-5_22

unfolding events, Reddit hosts a variety of subreddits on different topics, providing more detailed and contextually rich discussions [3]. Its extensive use as a data source spans applications such as predicting popularity, flagging hate speech, trolling identification, and user activity analysis [14]. While considerable attention has been devoted to the challenges of specific social network analytical methods, few studies have addressed the critical stages of data discovery, collection, and preparation [22]. This paper thus focuses on the issue of data veracity which is one of the primary challenges within the realm of data preparation [12]. The accuracy of data is paramount, as inaccuracies can distort analysis outcomes, leading to false discoveries and biases.

Conventional methodologies for data collection from Reddit typically entail either keyword-based post retrieval or targeting specific subreddit communities. Despite the initial filtration provided by these methods, the efficacy of these traditional strategies in ensuring data quality and relevance remains a subject of scrutiny and necessitates further refinement. Human labeling has often served as a method for post-annotation. This method has multiple drawbacks, including labor-intensive human labeling and the risk of bias, leading to inconsistent annotations and compromising data reliability [10]. Supervised content filtering algorithms offer a potential alternative to human labeling. While such algorithms have been developed for platforms like Twitter and Flickr, their applicability to Reddit data is limited. Moreover, the scarcity of pre-existing annotated datasets for Reddit increases the difficulty of training and validating these algorithms for effective post-filtering.

In this study, as outlined in Fig. 1, we address these gaps by proposing a novel semi-supervised domain adaptation network designed to automatically label Reddit posts using pre-existing human-labeled data from a related but distinct domain, namely Twitter (now known as X). We propose the generation of weak labels for Reddit by extracting information from Wikipedia pages associated with relevant subreddits using the YAKE algorithm [5]. Subsequently, we leverage the manually annotated Twitter dataset, TREC microblog [15] and combine it with weakly labeled Reddit posts to assign relevance labels within specific subreddit communities. Our approach utilizes a domain adversarial adaptation network to minimize domain discrepancy between Twitter and Reddit data. In the course of this work, we endeavor to address the following research questions:

- Can we automatically label Reddit data by utilizing annotated data from a different but related source (Twitter)?
- Can leveraging unlabeled or weakly labeled data from a related source enhance model performance across both labeled and unlabeled domains?
- What level of contextual information is necessary to effectively train the domain adaptation model?
- What is the most effective weak labeling method to complement the semi-supervised domain adaptation?

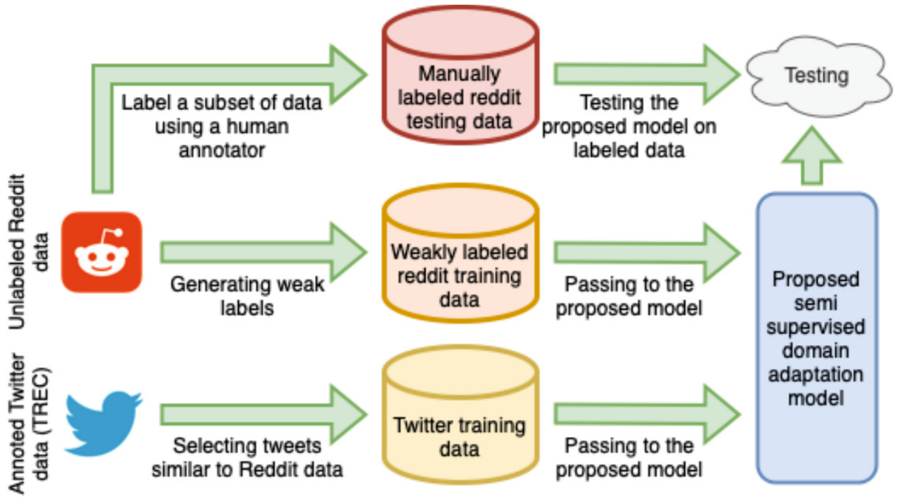
Unlabeled Reddit data
Annoted Twitter data (TREC)
Label a subset of data using a human annotator
Manually labeled reddit testing data
Testing the proposed model on labeled data
Testing
Generating weak labels
Weakly labeled reddit training data
Passing to the proposed model
Selecting tweets similar to Reddit data
Twitter training data
Passing to the proposed model
Proposed semi supervised domain adaptation model

Fig. 1. Study overview: Reddit data was collected and weak labels were generated. The model was trained on human-labeled Twitter data from the TREC microblog and weakly labeled Reddit data, while a human-annotated Reddit dataset was reserved for testing.

2 Related Work

Traditional supervised learning methods have been predominant in content-filtering research. For instance, techniques such as topic modeling [17], Support Vector Machines [13] and deep learning [1] have been applied to identify and filter out undesirable content. One study used BERT to find relevant comments on YouTube [18]. However, a key limitation of these approaches is their reliance on in-domain methods, often trained and tested on platforms other than Reddit. Reddit, unlike many other platforms, heavily relies on volunteer moderators to curate its content. Despite their efforts, the sheer volume of posts and the focus of moderators on flagging disruptive content can result in background chatter or irrelevant posts slipping through the cracks [19], thereby impacting the overall quality of the data available for research purposes. Moreover, the supervised approaches necessitate large manually labeled datasets for training, which are scarce and costly to obtain for platforms like Reddit. Consequently, alternative approaches are warranted to mitigate these challenges and enhance the effectiveness of content filtering on Reddit.

While some studies have explored cross-domain analysis for tasks such as fake news detection [21] and collaborative filtering [11], none have specifically addressed the challenge of filtering general irrelevant content on Reddit using labels generated from a different domain, such as Twitter. One study aimed to create user interest datasets from Twitter using Reddit posts [8]. However, the methodology relied solely on topic modeling and did not incorporate transfer learning or domain adaptation techniques for the task. In our study, we introduce a novel cross-domain strategy for content filtering on Reddit, harnessing data from Twitter and Wikipedia to automatically generate relevant labels. Our proposed framework combines large language models (LLMs), feed-forward neural networks, and domain adaptation adversarial networks. We draw from the recent findings that fine-tuning LLMs may not suffice when handling diverse data sources and domain adaptation techniques are needed to improve the model's ability to generalize across different domains [7]. By leveraging data from Twitter, our approach aims to improve the robustness and adaptability of the proposed automatic content filtering model on Reddit.

3 Datasets

Three different data sources are used in this study- 1) Reddit, 2) Twitter and 3) Wikipedia. Reddit data is furthermore divided into a labeled dataset that is used for testing and a weakly labeled dataset that is used for training. Each of the datasets is discussed further in the subsections.

3.1 Reddit Dataset

In this study, we adopt technology as a comprehensive thematic umbrella under which we select subreddits manually. We gathered posts from 50 diverse technology-based subreddits spanning the entirety of the year 2022 using Pushshift API [2]. These subreddits encompassed a wide array of categories, including but not limited to apps, web browsers, and gaming consoles. Each post was associated with its originating subreddit which we interpret as its corresponding 'topic'. The selection of subreddits was deliberately varied to ensure that our machine learning model remained uninfluenced by any potential writing biases inherent to specific topics or communities.

Subsequently, to create a labeled dataset, we randomly sampled 20 posts from each subreddit. The annotator was tasked with assessing the relevance of each post to a given technology topic, marking posts as '1' if deemed relevant and '0' if judged irrelevant. This process results in a binary classification task, where the goal is to determine the relevance of posts to the specified technology topic. By employing this rigorous manual annotation procedure, we aimed to create a robust dataset that can act as ground truth for the model evaluation. In total, we created 1000 labeled rows containing samples from 50 different subreddits from Reddit. Within our dataset, we have identified 73.6% of posts labeled as 'relevant' and 26.4% of posts labeled as 'irrelevant'. Note that the labeled dataset is hidden from the model while training and only used while reporting testing accuracy. Therefore, the model does not have access to any manually labeled examples from Reddit data sources.

To construct a comprehensive training dataset, we randomly sampled 100 posts from each of the 50 subreddits. Unlike the labeled dataset, these posts were not subjected to manual annotation. Instead, we employed a methodology detailed in Sect. 4.1 to generate weak labels for these posts. This involved leveraging information posted on Wikipedia about a technology to assign labels to the posts based on certain criteria. By implementing this strategy, we aimed to produce a training dataset from a target domain with a diverse range of examples while ensuring an efficient and quick weak label generation process.

3.2 TREC Microblog Dataset (Twitter)

The TREC (Text Retrieval Conference) microblog dataset is a benchmark dataset commonly used in information retrieval research. It consists of posts (tweets) collected over specified periods of time from Twitter. These tweets have been manually annotated as relevant or irrelevant to specific topics. In our proposed methodology, the dataset serves as a reliable source of strong labels. While the dataset contains data from years 2011 to 2015, to leverage transfer learning effectively, we opted to train our model using the labeled dataset spanning the years 2014 and 2015. The rationale behind selecting these specific years stems from the inclusion of a concise description for each topic which was missing prior to 2014. It is like the subreddit descriptions and aids in aligning the training data with the context of our Reddit training and testing dataset.

In our transfer learning approach, we strategically selected instances by employing a cosine similarity metric. This metric enabled us to identify examples from the TREC microblog dataset that exhibited more similarity to our Reddit dataset. By doing so, we aimed to ensure that the labeled training instances closely resembled our unlabeled data, thereby facilitating a more effective knowledge transfer. For both positive and negative samples, we choose examples from each class that have a cosine similarity of 0.7 and higher from the labeled dataset. In the end, we have 4959 examples with 46% 'relevant' and 54% 'irrelevant' samples.

It is essential to recognize that Twitter and Reddit, though both categorized as social media platforms, represent distinct data sources with unique characteristics and the nature of the data distribution varies significantly between them. Twitter imposes a character limit of 140 characters per tweet, promoting brevity and conciseness in communication. Conversely, Reddit allows for more extensive and descriptive posts, without any character limitations. This fundamental difference poses challenges in directly applying methodologies across both platforms. Although cosine similarity serves as a valuable metric for instance selection in tweets, effectively bridging the gap between Twitter and Reddit data requires a more nuanced approach. Domain adaptation emerges as a critical strategy in matching mean embeddings of the two distributions. Domain adaptation techniques aim to adapt models trained on one domain (e.g., Twitter) to perform well on target domain (e.g., Reddit).

3.3 Wikipedia

Our method for generating weak labels for Reddit posts leverages the text available on Wikipedia. For each technology based subreddit, we used the Wikipedia API to identify the most closely related Wikipedia page. Then we performed web scraping techniques to download the textual information from the chosen page thereby collecting rich and diverse contextual information about the technology.

4 Methodology

The primary objective of our research is to develop a predictive model capable of determining the relevance of Reddit posts to the topic of discussion that minimizing human intervention through a semi-supervised domain transfer capable of learning transferable features between Reddit and Twitter data sources. This methodology begins with the generation of weak labels for Reddit posts. Subsequently, we adopt a model fusion strategy that integrates the power of two large language models within a neural network framework. Specifically, we feed the topic and post inputs to the BERT and BERTweet models, respectively, to extract contextual embeddings. These embeddings are then further processed through a neural network architecture that incorporate domain adversarial training of neural networks to enhance model generalization across Reddit and Twitter domains. The methodology is further elaborated in the following subsections.

4.1 Weak Label Generation

YAKE (Yet Another Keyword Extractor) is an algorithm used for keyword generation and extraction from text [5]. To facilitate our methodology, we utilized the Wikipedia API to access and retrieve the textual content of the Wikipedia page corresponding to each technology of interest. Employing the YAKE algorithm, we conducted keyword extraction from the obtained Wikipedia text, identifying the top 50 keywords associated with each subreddit's respective technology. Consequently, we established distinct keyword lists for each of the 50 subreddits, encapsulating the most relevant terms. Next, we proceeded to sample 100 random unlabeled posts from each subreddit. If a post contained one or more keywords from its associated keyword list, it was attributed a weak label of '1'; conversely, if no such keywords were found, the post received a weak label of '0'. This approach enabled us to assign preliminary weak labels to the posts, facilitating the subsequent training process of our predictive model. After this process, we got 68% positive labels and 32% negative labels.

4.2 Model Architecture

Upon completing the creation of the weakly labeled Reddit dataset and the TREC dataset, our input is comprised of two distinct components: the topic and the post. In the case of Reddit data, the topic encompasses the subreddit name and its accompanying description, which are separated by a '[SEP]' token. Conversely, for the TREC dataset, the topic comprises the topic name and its associated description, as provided within the dataset itself. For Reddit data, the posts correspond to individual Reddit submissions, while for the TREC dataset, they represent tweets extracted from Twitter. Subsequently, the input data are fed into the proposed model, as illustrated in Fig. 2. When processing input data consisting of topics and posts, we adopt a tailored approach that capitalizes on the strengths of different language models to handle the distinct characteristics of each component. Topics, being more formally structured in language, are fed into a BERT model. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language model known for its ability to capture complex linguistic patterns and relationships within text data [6]. On the other hand, posts-particularly those sourced from Twitter-exhibit unique characteristics such as abbreviations, slang, hashtags, and mentions, which are pervasive in the informal language of tweets. To effectively process such content, we utilize a BERTweet model [16]. BERTweet is specifically pre-trained on a vast corpus of Twitter data, enabling it to understand and encode Twitter text more adeptly than traditional BERT models. We take embeddings of the last hidden layers from both models and reshape it to form a 1D input. Once the embeddings are generated from both the BERT and BERTweet models, they are combined and fed into a feed forward neural network. This neural network architecture integrates the information from both embeddings and generates a final relevance prediction output. By harnessing the strengths of both BERT and BERTweet embeddings, our model achieves a more comprehensive understanding of the input data, leading to more accurate and insightful predictions regarding relevance. Note that for both transformer models, only the last four layers are fine-tuned while remaining layers are frozen during training to reduce training time.

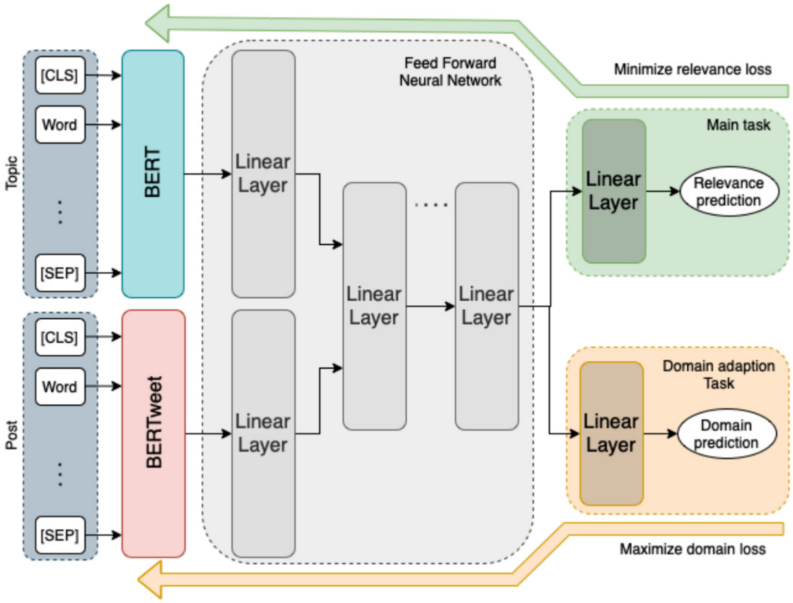

Fig. 2. Proposed methodology for semi-supervised domain adaptation model: Input (topics and posts) undergo processing via BERT, BERTweet, and a feed-forward neural network to predict relevance labels. The model minimizes relevance loss while maximizing domain loss, enabling effective domain adaptation.

Next, we turn our attention to domain adaptation. By bridging the gap between Twitter and Reddit, our model gains robustness and effectiveness across both platforms. This ensures its ability to accurately analyze and predict outcomes on text data sourced from diverse origins. In our approach, we draw inspiration from the DANN (Domain-Adversarial Neural Network) concept [9]. We incorporate a secondary classification task within our model- a linear classification task specifically designed for domain classification. In this setup, the model is tasked with determining whether a given instance originates from Twitter or Reddit. However, contrary to conventional classification objectives, rather than minimizing the loss associated with domain classification, we seek to maximize it. This deliberate inversion of the loss function ensures that the model refrains from learning domain-specific features that could potentially introduce bias or hinder generalization across domains. By prioritizing the maximization of domain classification loss, we make sure that the model focuses solely on extracting common characteristics across domains relevant to the primary learning task. The learning objective of the model is to minimize the total loss computed as-

$$Relevance\_loss = \sum_{t=1}^{N}(NLL_{loss}(\hat{y}_s, y_s) + \beta NLL_{loss}(\hat{y}_t, y_t)) \tag{1}$$

$$Domain\_loss = \sum_{t=1}^{N}(NLL_{loss}(\hat{l_s}, 0) + NLL_{loss}(\hat{l_t}, 1)) \tag{2}$$

$$Total\_loss = Relevance\_loss - \gamma Domain\_loss \tag{3}$$

where $\hat{y}_s$ and y_s are the predictions and ground truth for a given batch from source data (Twitter) respectively. $\hat{y}_t$ and y_t are the predictions and ground truth for a given batch from target data (Reddit) respectively. $\hat{l_s}$ and $\hat{l_t}$ are predicted domain labels for source and target domain instances respectively. N is the batch size, γ is domain transfer loss weightage and β is weak label loss weightage.

4.3 Experimental Setup

We hypothesize that it is feasible to generate relevancy labels for Reddit posts by performing semi-supervised learning from a relevant domain under different distribution (Twitter). Both the TREC-labeled dataset and the Reddit weak-labeled dataset are sent as input to the framework. Through a process of hyper-parameter fine-tuning using brute force method, we determined that a learning rate of 2e-5, a domain transfer loss weightage of 0.7, and a weak label weightage of 0.4 yielded optimal results for our model training procedure. The dropout rate is 0.2, and the batch size of 50. For BERT, a token length of 64 and for BERTweet, 128 token length yielded the best result. We used the Negative Log-Likelihood loss function for both prediction and domain classification and ADAM optimizer during training.

5 Results

We conducted extensive experiments to evaluate the performance of the proposed model and compare it to baselines. In this section, we address the research questions described in Subsects. 5.1 to 5.4 to evaluate the performance of the model. Macro F-1 score, and macro accuracy were used to evaluate the model's performance as the testing data is imbalanced.

5.1 Can We Automatically Label Reddit Data by Utilizing an Annotated Data from a Different but Related Source (Twitter)?

Through a series of meticulously designed experiments, we aim to substantiate our hypothesis and contrast alternative approaches against our proposed model, comprising human-annotated labels from Twitter, weakly labeled Reddit data based on presence of relevant terms determined using Wikipedia, and a domain adaptation network, against alternative approaches. For the baselines, two different configurations were employed: (1) utilizing the same architecture as the proposed model but with varied inputs, and (2) modifying the architecture to exclude the domain prediction task. As depicted in Table 1, our model's macro accuracy and macro F1 score are compared with various baseline models. Notably, models trained solely on Twitter data and weakly labeled Reddit datasets exhibit inadequate performance in accurately predicting relevancy labels for our labeled Reddit dataset. Similarly, a model trained on both Twitter data and weakly labeled Reddit data, albeit lacking domain adaptation, yields an accuracy of 41%. Furthermore, the incorporation of domain adaptation into a model trained on Twitter and unlabeled Reddit data only marginally improves accuracy to 53%, falling short of optimal performance. Conversely, our proposed model, integrating the features, increases the accuracy by 20%. It achieved an

Table 1. Performance of the proposed method and four alternatives tested on the target domain (Reddit). Two configurations were employed for baselines: (1) same architecture as the proposed model but with varied inputs, and (2) modified architecture that excludes the domain prediction task.

Models	Accuracy	F1 score
Model trained on Twitter without domain adaptation	31%	0.22
Model trained on Reddit weak labels without domain adaptation	52%	0.48
Model trained on Twitter and unlabeled Reddit data with domain adaptation	41%	0.39
Model trained on Twitter and Reddit weak labels without domain adaptation	53%	0.46
Proposed model	**73%**	**0.77**

impressive accuracy of 73% and F1 score of 0.77. In summary, these baseline experiments highlight the enhanced performance attributable to the inclusion of weak labels for Reddit, robust human-annotated labels for Twitter, and domain adaptation. Our findings support the efficacy of utilizing pre-existing annotated data from a related albeit distinct domain, Twitter, to achieve competitive labeling results for Reddit.

5.2 Can Leveraging Unlabeled or Weakly Labeled Data from a Related Source Enhance Model Performance Across both Labeled and Unlabeled Domains?

Incorporating annotated data from the source domain of Twitter has proven advantageous in generating labels for the target domain of Reddit. Building upon this observation, our investigation delves deeper into the potential benefits of integrating unlabeled or weakly labeled data from the target domain to enhance the model's generalization capacity or specifically, enhances the model's performance on source domain, Twitter. Our analysis, presented in Table 2, denotes the performance gains achieved by our model when compared against various baseline approaches on Twitter data. Notably, the inclusion of Reddit data in any form, whether unlabeled or weakly labeled, consistently amplifies the model's performance by a minimum of 3%. This empirical evidence highlights the efficacy of domain adaptation techniques in facilitating the model's comprehension of domain-invariant features. Specifically, the integration of unlabeled Reddit data, coupled with domain adaptation, elevates accuracy from 71% to 79%, showing the substantial impact of domain-aware methodologies. Moreover, our proposed model, incorporating weakly labeled Reddit data alongside domain adaptation, attains optimal performance metrics, boasting an impressive accuracy and F1 score of 0.9 each. This substantiates the effectiveness and robustness of our approach in leveraging both human - labeled and human - unlabeled data sources to achieve superior performance across domains.

Table 2. Performance of the proposed method and three alternatives tested on the source domain (Twitter).

Models	Accuracy	F1 score
Model trained on Twitter without domain adaptation	71%	0.76
Model trained on Twitter and unlabeled Reddit data with domain adaptation	79%	0.78
Model trained on Twitter and Reddit weak labels without domain adaptation	74%	0.74
Proposed model	**90%**	**0.90**

5.3 What Level of Contextual Information Is Necessary to Effectively Train the Domain Adaptation Model?

Our research methodology involves a comprehensive examination of our input data, which is comprised of two distinct components: posts and topics. For Reddit, posts consist of user-generated content within the platform, while for Twitter, they correspond to tweets. Regarding topics, for Reddit, we utilize both the subreddit name and its associated description, whereas for Twitter, we leverage the topic and description provided by the TREC dataset. The critical inquiry lies in determining the optimal amount of information required for effectively characterizing the topic component of our input. To address this, we investigate models that are furnished with varying degrees of topic information: one utilizing solely the subreddit name, another incorporating both, the subreddit name and description (as proposed by our methodology), and a third model augmented with additional context from Wikipedia pages related to the discussed technology. The outcomes of these experiments are detailed in Table 3. Notably, our findings reveal that providing the subreddit name with its corresponding description yields a notable improvement in accuracy, with a 2% increase compared to models reliant solely on the subreddit name. However, intriguingly, the incorporation of Wikipedia knowledge fails to confer any noticeable performance enhancement, indicating that the additional contextual information does not significantly contribute to the model's predictive capabilities. These results show the importance of strategic feature selection in optimizing model performance and highlight the limited utility of additional information in certain contexts.

Table 3. Performance of the proposed method trained in different topic inputs for Reddit.

Contextual information provided for the topic input	Accuracy	F1 score
Subreddit name	71%	0.74
Subreddit name + description	**73%**	**0.75**
Subreddit name + description + Wikipedia page	72%	0.75

5.4 What Is the Most Effective Weak Labeling Method to Complement the Semi-supervised Domain Adaptation?

An integral aspect of our model lies in the implementation of the weak labeling method. Our proposed approach employs the YAKE algorithm to extract keywords from Wikipedia page relevant to the topic, subsequently assigning binary labels (1 for relevant, 0 for irrelevant) to Reddit posts based on the presence or absence of these keywords. To comprehensively assess the effectiveness of our method, we conducted an analysis against alternative approaches, each designed to determine the relevance of Reddit posts through distinct means.

- **Subreddit Name Criterion:** In this method, we designated a post as relevant if it contained the name of the corresponding subreddit; otherwise, it was labeled as irrelevant. This simplistic criterion serves as a baseline for comparison.
- **Wikipedia Hyperlink Criterion:** Here, we leveraged keywords extracted from Wikipedia pages with hyperlinks. Posts containing any of these keywords were labeled as relevant, while others were deemed irrelevant. Notably, this method boasts a lower time complexity compared to the YAKE algorithm.
- **KeyBERT-based Criterion:** Employing the KeyBERT model [20], we extracted keywords from the Wikipedia page text and utilized them to assess the relevance of Reddit posts. KeyBERT, a transformer-based model for keyword extraction, utilizes BERT embeddings to identify significant keywords within text. However, it is important to note that this method entails a higher time complexity relative to the YAKE algorithm.

For YAKE and KeyBERT method, the top 50 keywords were found for each topic. For hyperlinks criterion, all the keywords from a topic's Wikipedia page with hyperlinks are stored in a list and length of the keyword lists can vary from topic to topic. Note that this is not a comprehensive list of all possible weak labelling methods and we will explore other methods in future work.

Results shown in Table 4 provide evidence that the YAKE algorithm-based weak labeling method outperforms the established baselines by a minimum margin of 3% in accuracy. Specifically, the superior performance of YAKE can be attributed to its specialized optimization for identifying important terms within textual data. Unlike simplistic criteria such as the subreddit name criterion, which solely relies on the presence of specific terms, YAKE leverages a better understanding of context to judge relevance. For instance, in scenarios where a post originates from a subreddit such as 'Adsense', which pertains to discussions on Google Ads, YAKE's capability to recognize the contextual significance of terms allows it to label posts containing 'Google' as potentially relevant, even in the absence of explicit subreddit references like 'Adsense'. Conversely, the subreddit name criterion may overlook such posts, erroneously classifying them as irrelevant. Similarly, when compared to methods reliant on hyperlink analysis, YAKE demonstrates a more sophisticated approach to keyword identification. Lastly, while both YAKE and KeyBERT offer advanced methodologies for keyword extraction, a potential advantage of the YAKE-based method lies in its ability to identifying key phrases that encapsulate the essence of the document.

Table 4. Performance of the proposed method trained in different weak labels setting.

Weak label generation method	Accuracy	F1 score
Name of subreddit	69%	0.70
Hyperlinks from Wikipedia page	64%	0.61
Yake based keywords from Wikipedia page	**73%**	**0.77**
KeyBERT based keywords from Wikipedia page	70%	0.66

In contrast, KeyBERT primarily focuses on generating embeddings and representations of text, thereby potentially compromising the precision of keyword selection. This inherent difference in focus can lead to the superiority of the YAKE algorithm and make it better suited for our purposes.

6 Conclusion

In this study, our primary objective was to generate relevancy labels for Reddit posts, thereby facilitating binary classification to decide whether a given post aligns with a specified topic. Due to the scarcity of adequately manually-annotated data for this task, we embarked on an exploration of semi-supervised domain adaptation techniques, leveraging data from a related but distributionally different source: Twitter. Using the TREC microblog dataset for human-labeled data, coupled with the YAKE algorithm for weakly labeling the Reddit dataset based on presence of keywords extracted from related Wikipedia pages, we subsequently augmented domain adversarial adaptation networks into our approach. We have demonstrated that our proposed model is capable of accurately predicting relevancy labels for Reddit posts, achieving an accuracy of 73% and an F1 score of 0.77, a +20% improvement when compared to baseline models.

Furthermore, our investigation revealed that our proposed architecture not only excels in the target domain of Reddit but also enhances the model's generalization capability, leading to improved performance in the source domain of Twitter. This observation shows the efficacy and versatility of our approach in adapting to diverse data distributions and domains. Additionally, we demonstrated the benefits of the proposed weak labeling method and showed the significance of conceptual information pertaining to the topic component of input data hence contributing to the broader discourse on weak labeling methodologies in text classification tasks.

Our study utilized BERTweet for processing posts from Reddit which technically was not designed for this purpose. Future research could explore alternative methodologies, such as leveraging GPT for input processing or training long transformers from scratch. Furthermore, it is important to note that our evaluation dataset was labeled by a single annotator, which may introduce biases in the annotations and future works can focus on overcoming this shortcoming. Moving forward, future works can delve into comparing the proposed architecture with other relevance filtering models and the development of a generalized model capable of integrating labeled and unlabeled data from various sources, including but not limited to Facebook, blogs, and news sources. Such a universal model would possess the versatility to predict relevancy labels for text from any source, thereby enhancing its applicability and utility across diverse domains. Lastly, future investigations can study into alternative weak labeling and domain adaptation methods that can enhance model performance.

Acknowledgements. Research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Also supported, in part by the Temple University office of the Vice President for Research 2022 Catalytic Collaborative Research Initiative Program, AI & ML Focus Area.

References

1. Alharthi, R., Alhothali, A., Moria, K.: A real-time deep-learning approach for filtering Arabic low-quality content and accounts on Twitter. Inf. Syst. **99**, 101740 (2021)
2. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J.: The Pushshift reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 830–839 (2020)
3. Bonifazi, G., Corradini, E., Ursino, D., Virgili, L.: Modeling, evaluating, and applying the eWoM power of reddit posts. Big Data Cognit. Comput. **7**(1), 47 (2023)
4. Camacho, D., Panizo-LLedot, A., Bello-Orgaz, G., Gonzalez-Pardo, A., Cambria, E.: The four dimensions of social network analysis: an overview of research methods, applications, and software tools. Inf. Fus. **63**, 88–120 (2020)
5. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., Jatowt, A.: YAKE! Keyword extraction from single documents using multiple local features. Inf. Sci. **509**, 257–289 (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Du, C., Sun, H., Wang, J., Qi, Q., Liao, J.: Adversarial and domain-aware BERT for cross-domain sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4019–4028 (2020)
8. Fiallos, A., Jimenes, K.: Using reddit data for multi-label text classification of twitter users interests. In: 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG), pp. 324–327. IEEE (2019)
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(59), 1–35 (2016)
10. Geiger, R.S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., Huang, J.: Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 325–336 (2020)
11. Hu, G., Zhang, Y., Yang, Q.: Transfer meets hybrid: a synthetic approach for cross-domain collaborative filtering with text. In: The World Wide Web Conference, pp. 2822–2829 (2019)
12. Kepner, J., et al.: Computing on masked data: a high-performance method for improving big data veracity. In: 2014 IEEE High Performance Extreme Computing Conference (HPEC), pp. 1–6. IEEE (2014)

13. Kumaresamoorthy, N., Firdhous, M.: An approach of filtering the content of posts in social media. In: 2018 3rd International Conference on Information Technology Research (ICITR), pp. 1–6. IEEE (2018)
14. Medvedev, A.N., Lambiotte, R., Delvenne, J.C.: The anatomy of Reddit: an overview of academic research. In: Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches, vol. 10, pp. 183–204 (2019)
15. National Institute of Standards and Technology (NIST): TREC Microblog Track (2024). https://trec.nist.gov/data/microblog.html. Accessed 21 Feb 2024
16. Nguyen, D.Q., Vu, T., Nguyen, A.T.: BERTweet: a pre-trained language model for English Tweets. arXiv preprint arXiv:2005.10200 (2020)
17. Nutakki, G.C., Nasraoui, O.: Compartmentalized adaptive topic mining on social media streams. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 992–997. IEEE (2016)
18. Ramamonjisoa, D., Ikuma, H., Murakami, R.: Filtering relevant comments in social media using deep learning. In: 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 335–340. IEEE (2022)
19. Seering, J., Wang, T., Yoon, J., Kaufman, G.: Moderator engagement and community development in the age of algorithms. New Media Soc. **21**(7), 1417–1443 (2019)
20. Sharma, P., Li, Y.: Self-supervised contextual keyword and keyphrase retrieval with self-labelling (2019)
21. Silva, A., Luo, L., Karunasekera, S., Leckie, C.: Embracing domain differences in fake news: cross-domain fake news detection using multi-modal data. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35, pp. 557–565 (2021)
22. Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C.: Social media analytics-challenges in topic discovery, data collection, and data preparation. Int. J. Inf. Manage. **39**, 156–168 (2018)