AVA-TMP: A Human-in-the-Loop Multi-layer Dynamic Topic Modeling Pipeline

Viseth Sean, Padideh Danaee, Yang Yang, Hakan Kardes
Alignment Health, Orange, CA
{vsean, pdanaee, yyang, hkardes}@ahcusa.com

Abstract

A phone call is still one of the primary preferred channels for seniors to express their needs, ask questions, and inform potential problems to their health insurance plans. Alignment Health is a next-generation, consumer-centric organization that is providing a variety of Medicare Advantage Products for seniors. We combine our proprietary technology platform, AVA, and our high-touch clinical model to provide seniors with care as it should be: high quality, low cost, and accompanied by a vastly improved consumer experience. Our members have the ability to connect with our member services and concierge teams 24/7 for a wide variety of ever-changing reasons through different channels, such as phone, email, and messages. We strive to provide an excellent member experience and ensure our members are getting the help and information they need at every touch — ideally, even before they reach us. This requires ongoing monitoring of reasons for contacting us, ensuring agents are equipped with the right tools and information to serve members, and coming up with proactive strategies to eliminate the need for the call when possible.

We developed an NLP-based dynamic call reason tagging and reporting pipeline with an optimized human-in-the-loop approach to enable accurate call reason reporting and monitoring with the ability to see high-level trends as well as drill down into more granular sub-reasons. Our system produces 96.4% precision and 30%-50% better recall in tagging calls with proper reasons. We have also consistently achieved a 60+ Net Promoter Score (NPS) score, which illustrates high consumer satisfaction.

1 Introduction

As a consumer-centered healthcare company, we provide our members with experienced member services and concierge teams through our contact center that are available around the clock for a wide variety of inquiries or potential problems. It is crucial for us to monitor the typical causes of why members contact us so that we proactively address the problems or help our members need through their preferred channel even before they reach out to us. To this end, during each call, member service agents take detailed notes on what is discussed during the call, the reason for the call, and any actions taken. They also tag each call with one or more call reason categories from a pre-defined list that was initially built by member experience supervisors to capture call reasons in a more structured manner. There are several drawbacks and limitations to this approach:

- High number of calls that are documented as "General FAQ" or "Other" call reason category due to lack of precise reason category representing the call
- Labor-intensive and not scalable processes to keep pre-defined call reason categories with corresponding subcategories manually up-to-date with high quality while member needs and potential call reasons are continuously changing
- Incomplete and inaccurate call reason reporting

In this paper, we present a novel NLP-based multi-layer dynamic topic modeling pipeline, AVA-TMP, with an effective human-in-the-loop approach that leverages subject matter experts. It enables highly accurate and timely call reason logging, monitoring, and reporting, as well as increased first-time resolution rates. Our pipeline produces a high-quality call reasons list with sub-reason drill-downs and automatically identifies newly emerging high-quality topics eliminating the need for labor-intensive evaluation of high-volume call notes by the customer service supervisors. Our pipeline also increases the efficiency of customer service agents by automatically suggesting a ranked list of relevant topics to tag as

Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances), pages 81 - 87
December 8, 2022 ©2022 Association for Computational Linguistics

agents take call notes. The presented pipeline is not limited to inbound calls, and it is applicable to omnichannel communications such as emails, messages, and chats. In the remainder of the paper, we first present the related work in section 2. Next, we describe our approach and evaluate its performance using several real-world datasets in section 3. Then, we evaluate the performance of our algorithm. Finally, we conclude in section 4.

2 Related Work

It has been a great stride in Natural Language Processing (NLP) advancement in recent years for learning, understanding, and producing human language content in more efficient and scalable ways (Hirschberg and Manning, 2015), such as text summarization (Widyassari et al., 2022), name entity recognition (Jiang et al., 2016), sentimental analysis (Bhavitha et al., 2017), text classification (Bhavani and Kumar, 2021), and topic modeling (Sandhiya et al., 2022).

Topic modeling is used to automatically identify the themes, i.e., topics, in unstructured text datasets (Blei et al., 2003b; Boyd-Graber et al., 2017), especially when there is a large volume of document collections and not enough time. These machine-generated topics are then used for reporting and tracking purposes. The traditional topic modeling pipelines (Chaney and Blei, 2012; Gardner et al., 2010; Eisenstein et al., 2012) do not allow the users to refine, clean, or personalize the model-generated topics hence lack an understanding of the end-user needs (Smith et al., 2018). The absence of an interactive and human-centered approach to refine topics and adjust data processing results in bias, noise, and unexpectedly poor real-world model performance. It is crucial to have an efficient and effective level of human contribution with AI to ensure satisfactory model performance.

Previous works have utilized topic modeling techniques with some level of human intervention to address customer needs. For example, a study by Agudelo and Manuel used Latent Dirichlet Allocation (LDA) (Blei et al., 2003b) based topic modeling approach to identify topics of inbound call transcripts for creating a better Interactive Voice Response (IVR) routing option so that customers are routed to the right agents (Agudelo and Manuel, 2019). However, they only provide the most important words to the end users and not the topic. The users then need to manually identify the topic associated with the words presented by the model which can be inconsistent among different users and adds an additional burden to their workflow. This process also results in assigning only one topic to a call which might not be a true representation of customer needs as one call might have multiple reasons and resolutions.

Another study by Chen and Wang utilized the LDA topic modeling technique to extract topics from chat transcripts between librarians and students in order to identify needs, provide help and allocate resources accordingly (Chen and Wang, 2019). However, this work lacks defining a qualitative performance measurement. Also, they assessed the quality of the topics intuitively with visualizations.

Besides topic modeling methods from machine learning, there are also qualitative analytic methods, such as grounded theory methodologies (Charmaz, 2006; Corbin and Strauss, 2008), for identifying topics in text datasets. Previous work by Baumer et al. focused on the comparison between grounded theory and topic modeling (Baumer et al., 2017). The authors found that the results of the two methods exhibited a degree of alignment in which many of the patterns found in the grounded theory were also represented in the topic modeling results. However, grounded theory is time-consuming and resource-intensive. Therefore, it doesn't scale well with large datasets.

Different from these previous works, we implemented a novel NLP-based system with human-in-the-loop stages to accurately tag high-level call reasons (one or multiple) along with sub-category drill-downs in a scalable and timely manner for each call. We also measured the real-world performance of our approach and compared it with the baseline (manual labeling).

3 Methodology

Our comprehensive end-to-end topic modeling pipeline, AVA-TMP, is illustrated in Figure 1. With AVA-TMP, we enhance the traditional NLP topic modeling pipeline with human-in-the-loop stages to significantly improve reporting accuracy, create operational efficiencies and increase member satisfaction.

Standard high-level topic modeling steps include Problem definition, Data collection, Data preparation (tokenization, lemmatization, and stop-word removal), Modeling, Post-processing & model eval-

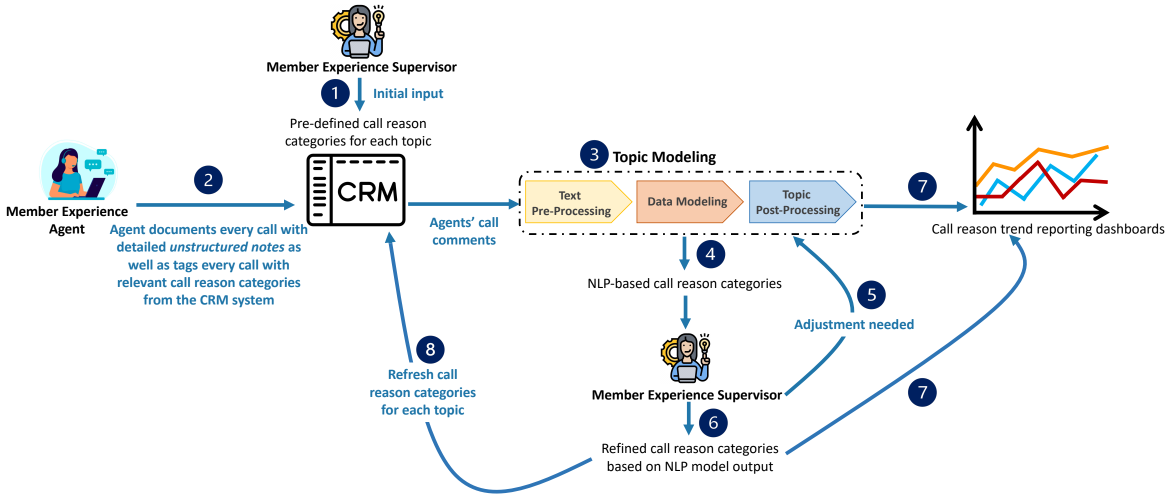
Member Experience Supervisor
1
Initial input
Pre-defined call reason categories for each topic
Member Experience Agent
2
Agent documents every call with detailed *unstructured notes* as well as tags every call with relevant call reason categories from the CRM system
CRM
Agents' call comments
3
Topic Modeling
Text Pre-Processing
Data Modeling
Topic Post-Processing
7
Call reason trend reporting dashboards
4
NLP-based call reason categories
5
Adjustment needed
Member Experience Supervisor
6
Refined call reason categories based on NLP model output
7
8
Refresh call reason categories for each topic

Figure 1: **Overview of AVA-TMP human-in-the-loop pipeline:** 1) Member experience supervisors create an initial list of call reasons for member experience agents to choose from. 2) Member experience agents take notes during calls and also select a reason why member calls from the initial list of reasons in 1). 3) We perform topic modeling on call notes to identify the reason(s) why members call us. 4) The NLP-based reasons are transferred to member experience supervisors for evaluation. 5) Member experience supervisors refine and adjust the newly generated list of reasons by NLP and send it back to step 3) if necessary. 6) Member experience supervisors finalize the new list of call reasons based on the NLP model. 7) We create a call reason trend reporting dashboard based on the new call reason list. 8) We refresh the call reason categories in the CRM system so member experience agents can properly assign reason tags to upcoming calls.

uation, and Model deployment. We altered and improved the standard topic modeling pipeline in three ways:

- First, instead of using call transcripts, which has its own drawbacks and limitations, we use call notes that are already curated by member service agents for documentation purposes. These call notes provide a more contextual and cleaner description of each call (Figure 1 - step 2).
- Second, our pipeline involves subject matter experts (SMEs) to refine call reason categories and sub-reasons to further validate and adjust topics identified by the model (Figure 1 - steps 5 and 6).
- Third, we introduce a multi-layer dynamic hierarchical topic modeling framework to identify hierarchical call reasons and sub-reasons in a flexible manner (Figure 2).

In the following subsections, we describe the major end-to-end steps from initial data collection to the final model deployment, and how they are being used.

3.1 Data Collection

For every member interaction, our member experience agents take notes summarizing what help or information the member needs and what action(s) they take to assist the member. These unstructured call notes provide more contextual information about the call and more standardized and cleaner input to our model compared to using call transcripts. For example, even though actual call conversations might happen in different languages, all call notes are captured in English. Furthermore, there might be incomplete sentences, inaccuracies due to the performance of voice-to-text translation, and other complexities to deal with in call transcripts.

Besides capturing call notes, agents also need to tag each call based on the pre-defined call reason categories. The call reason tagging in this step can help us with immediate reporting within the CRM system so that member experience supervisors can view real-time reports on the pre-defined call reasons as a stopgap measure and respond to the severity of each issue accordingly. However, keeping this list up-to-date is labor-intensive and requires member experience supervisors to manually go through large volume comments, investigate call reasons, and identify frequent emerging patterns. As we describe in section 3.4, our model eliminates the labor-intensive call note reviewing step and helps to keep the call reasons list continuously refreshed. These call reason tagging also help us to measure the real-world performance of our model on an ongoing basis as described in detail in section 3.5.

3.2 Data Preparation

We clean the unstructured text from agents' notes using various standard text pre-processing techniques before feeding it into our model. First, we remove special characters, punctuation, and stopwords from the text, and then convert them to lowercase. Next, the text is lemmatized to return inflected words to their root word. We do not use stemming as it does not provide any performance improvement, which is in line with what other researchers found previously (Schofield and Mimno, 2016). Beyond these common text pre-processing techniques, we clean certain unnecessary and repetitive patterns that do not add value to the call tagging process such as confirmed demographics, or courtesy of callers. The processed dataset is then fed into the model.

3.3 Topic Modeling

After preparing the data, we utilize commonly used topic modeling techniques, LDA (Blei et al., 2003b) and a BERT-based method, BERTopic (Grootendorst, 2022) to achieve the best performance.

To accurately identify the call reason categories and sub-reason drill-downs, we construct a multi-layer topic modeling framework (as illustrated in Figure 2). First, we run the topic modeling with all the call notes to identify the high-level call reasons. Next, we run separate models to extract sub-reasons for each high-level call reason using the subset of call notes associated with the corresponding high-level topic. We repeat this step for another layer to get a further drill-down of each sub-reason. This approach gives us the flexibility to set the different number of topics at each stage and fine-tune the quality of topics. For example, we can start with 50 topics to identify high-level reasons, then in the next layer we can use 10 topics for sub-reasons, and finally 5 topics in the third layer to identify sub-sub-reasons. Each topic generated by the topic model is accompanied by their corresponding top *n* keywords that explain what the topic is about. We are also able to set different thresholds at each layer to optimize the quality of keyword groups for each topic/sub-topic/sub-sub-topic.

Another benefit of this multi-layer approach is the ability to pick different "n" for "n-gram" construction for the keywords representing each topic. For example, high-level topics have up to tri-grams since they are representing high-level reasons while sub-reasons and sub-sub-reasons are configured to have longer n-grams. We spent significant effort tuning parameters for both LDA and BERTopic within this framework in order to obtain the best possible results.

We choose BERTopic as our final topic modeling technique and 50/10/5 topics for different layers of our multi-layer topic modeling framework as it produced better performance and significantly reduced the review time by SMEs in our use case. Please also note that our multi-layer hierarchical topic modeling framework allows us to have more flexibility and control to adjust various parameters and fine-tune the quality compared to built-in hierarchical topic models like hLDA (Blei et al., 2003a).

3.4 Subject Matter Expert Evaluation

The NLP-based call reason categories from the topic modeling step are then passed to the member experience team for review (Figure 1 - step 4). They start evaluating the list of topics, pinpointing potential issues, identifying bias in the data, and requesting certain refinements such as grouping certain topics into one category, eliminating poor quality and biased topics, and choosing a more user-friendly topic name. These steps help fine-tune the NLP model performance and improve the processes to better understand call trends. This feedback loop might be repeated a few times till desired performance is achieved (Figure 1 - steps 3 to 5).

3.5 Performance Evaluation

In this section, we present the experimental results for our human-in-the-loop topic modeling pipeline, AVA-TMP. There are different ways to measure the performance of the topic models (Hoyle et al., 2021; Chang et al., 2009). We use two different real-world datasets for our experiments to measure both precision and recall of our system. We also compare the performance of the system with the baseline, which is the previous agent tagging using the pre-defined call reasons list during the call prior to developing AVA-TMP.

The first dataset includes randomly selected 500 actual call notes. We run AVA-TMP and generated

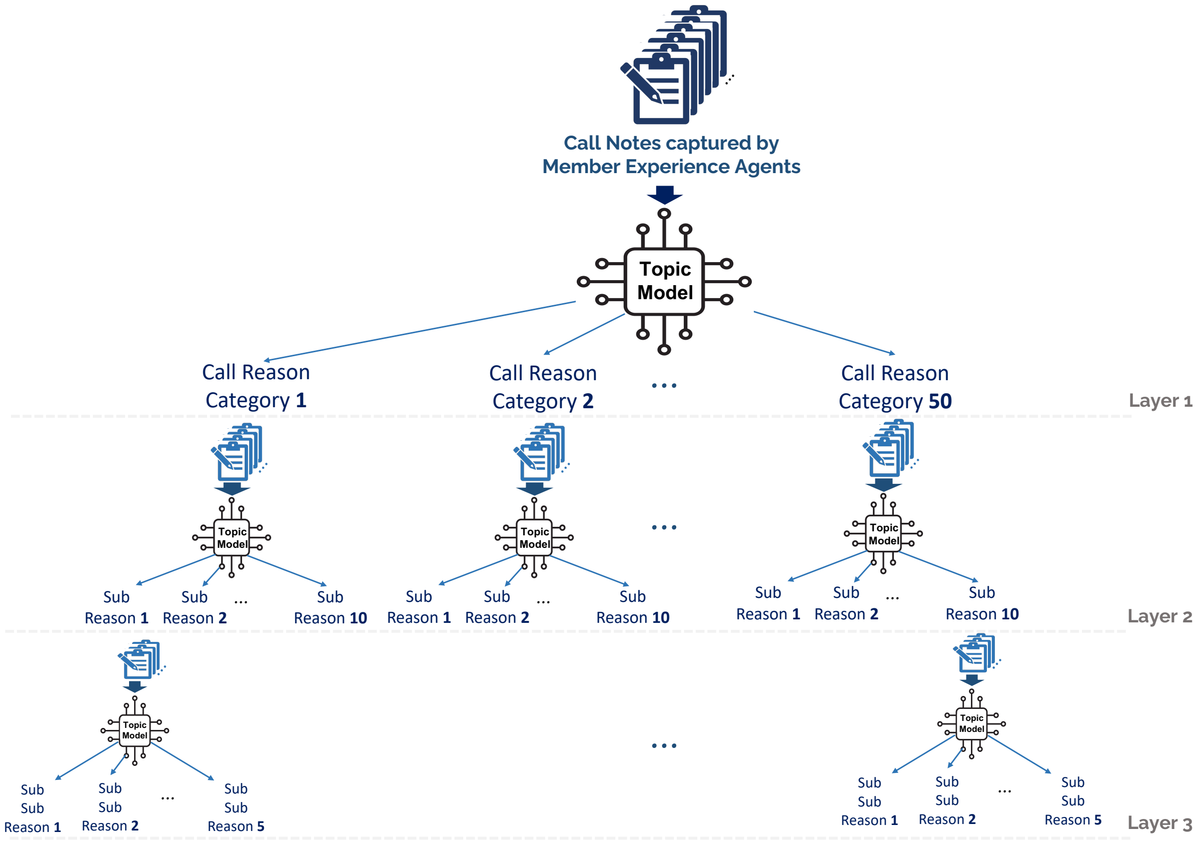

Figure 2: **Multi-layer Dynamic Hierarchical Topic Modeling**

all the call reasons for each call. As mentioned earlier, we already have baseline call reasons for these notes. Three independent SMEs evaluated both AVA-TMP output and baseline tags. SMEs reviewed and labeled each call note and corresponding reasons including all the drill-downs as True Positive (TP) or False Positive (FP). When there are conflicts among SMEs, we use the majority vote. We then computed the precision. AVA-TMP achieved a precision of 96.4% while the baseline precision was 23.4%.

Secondly, we use a much larger dataset to measure the recall improvement with the AVA-TMP system where high-level call reasons overlap with the baseline. We used 3 years of call notes history (1M+ calls) and compared the high-level call reason categories identified by AVA-TMP with the baseline. AVA-TMP achieves 30% to 50% higher recall than baseline depending on the high-level call reason categories. Please note that AVA-TMP identifies new call reason categories that don't exist in the baseline but we didn't include them in the recall analysis since the volume was negligible on a large scale.

3.6 Model Deployment

After generating refined call reason categories and sub-reason drill-downs (Figure 1 - step 6), we deploy our model to production to tag the upcoming calls on an ongoing basis for reporting and tracking purposes (Figure 1 - step 7). The pre-defined call reason categories are also updated in the CRM system accordingly for member experience agents to utilize (Figure 1 - steps 8).

The AVA-TMP is deployed as a real-time service. We run the model on the backend and provide suggested call reasons for tagging in a ranked order. Agents can then select one or multiple labels from the ranked topic list. This process aids the agents to minimize the time going through the entire list and decreases human error in the process.

3.7 Reporting & Tracking Call Trends

It is essential to continuously monitor the call trends to understand members' needs, proactively resolve emerging issues, and measure the member experience improvements. AVA-TMP provides timely and accurate call reasons after every call note is completed. We then feed all these into our reporting and tracking dashboard to monitor newly emerging issues and trends over time (Figure 1 - step 7).

4 Conclusion

In this paper, we demonstrate a novel topic modeling pipeline, AVA-TMP, that consists of multi-layer dynamic hierarchical topic modeling with the human-in-the-loop approach for call analytics. Our results illustrate having human-in-the-loop in an advanced NLP pipeline can optimize model performance, reduce manual tasks in the current workflows, and improve overall business outcomes while helping achieve high member satisfaction.

With this framework, we achieved significantly better performance using real-world datasets for evaluation. Alignment Health has also consistently achieved an overall NPS, which is a widely used metric to measure customer experience, of 60+ as a Medicare Advantage plan, compared to the industry average of 30 for healthcare insurance (Ian Luck, 2022; Alignment Healthcare, 2022).

AVA-TMP enables us to have an improved call reporting and tracking system. We can identify seasonal trends for demands, inquiries, or emerging issues and develop proactive and systemic plans for improvement. It also creates operational efficiencies by eliminating the need for various labor-intensive tasks such as i) having an up-to-date call reason list, ii) manual call reason tagging, and iii) manual ad-hoc reports.

Even though we illustrated the AVA-TMP pipeline on call notes, it is applicable to omnichannel communications such as calls, emails, messages, and chats. It can be also used on call transcripts in addition to the call notes for improved outcomes.

References

B. E. Llano Agudelo and Juan Manuel. 2019. Calls' topic recognition.

Alignment Healthcare. 2022. Alignment healthcare esg report. `https://www.alignmenthealth.com/Alignment/media/pdf/Alignment_ESGReport_080422_508.pdf`.

Eric Ps Baumer, David Mimno, Shion Guha, Emily Quan, and Geri Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68.

A. C. Bhavani and B.J Santhosh Kumar. 2021. A review of state art of text classification algorithms. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1484–1490.

Budeti Bhavitha, Anisha P. Rodrigues, and Niranjan N. Chiplunkar. 2017. Comparative study of machine learning techniques in sentimental analysis. *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 216–221.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*.

David M. Blei, A. Ng, and Michael I. Jordan. 2003b. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Jordan L. Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Found. Trends Inf. Retr.*, 11:143–296.

Allison Chaney and David Blei. 2012. Visualizing topic models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 419–422.

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Kathy Charmaz. 2006. Constructing grounded theory: A practical guide through qualitative analysis.

Xiaoju Chen and Huajin Wang. 2019. Automated chat transcript analysis using topic modeling for library reference services. *Proceedings of the Association for Information Science and Technology*, 56.

Juliet Corbin and Anselm Strauss. 2008. Basics of qualitative research: Techniques and procedures for developing grounded theory.

Jacob Eisenstein, Duen Horng Chau, Aniket Kittur, and Eric Xing. 2012. Topicviz: Interactive topic exploration in document collections. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2177–2182.

Matthew J Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2010. The topic browser: An interactive tool for browsing topic models. In *Nips workshop on challenges of data visualization*, volume 2, page 2. Whistler Canada.

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Julia Hirschberg and Christopher D. Manning. 2015. Advances in natural language processing. *Science*, 349:261 – 266.

Alexander Miserlis Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan L. Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken?: The incoherence of coherence. In *NeurIPS*.

Ian Luck. 2022. 25 insurance nps scores for 2022 + nps in insurance guide. https://customergauge.com/benchmarks/blog/nps-insurance-industry-net-promoter-scores.

Ridong Jiang, Rafael E. Banchs, and Haizhou Li. 2016. Evaluating and combining name entity recognition systems. In *NEWS@ACM*.

Ramesh Sandhiya, A. M. Boopika, M. K. Akshatha, S. V. Swetha, and N. M. Hariharan. 2022. A review of topic modeling and its application. *Handbook of Intelligent Computing and Optimization for Sustainable Development*.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*, pages 293–304.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. 2022. Review of automatic text summarization techniques & methods. *J. King Saud Univ. Comput. Inf. Sci.*, 34:1029–1046.