Use of a Fine-tuned, Bidirectional Transformer Large Language OEvidera PPD **MSR50** Model to Classify the Patient and Caregiver Voice through Their Social Media Health Posts: An Example in Non-small Cell Lung Cancer

Rai AK¹, Ikoro V², Berger A³

¹Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific, Overland Park, KS, USA; ²Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, Hammersmith, London, UK; ³Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific, Boston, MA, USA

Background

- Non-small-cell lung cancer (NSCLC) comprises 85% of lung cancer cases.¹ It is more resistant to treatment than small-cell variants and is associated with substantial negative emotional and physical impacts.
- "Traditional" sources of real-world data (RWD), such as healthcare claims and electronic medical records, do not typically contain large amounts of data on these emotional and physical impacts, which has unfortunately resulted in "muting" the voice of the patient (and their caregivers) on all aspects of disease and treatments therefor.
- Patient online forums, a relatively recent phenomenon, offer real-time insights into disease-related experiences of, and impacts to, patients and their caregivers and the FDA endorses use of such RWD to inform medical development and healthcare policies.^{2,3} While promising, accurate classification of posts from patients and caregivers is essential to obtain targeted insights; bidirectional encoder representations from transformers (BERT), a transformer model, may be able to improve this process by recognizing subtle language differences and providing more precise classification.

Results

- A total of 39,686 posts related to NSCLC were collected from seven patient forum sites. These posts were drawn from various forums where patients and caregivers share their experiences with NSCLC.
- The BERT model was used to predict the voice class of a total of 39,686 posts from seven forum sites. Of these, 15,632 posts were classified as "Other," 15,210 as "Patient," and 8,844 as "Caregiver" (Figure 3).
- Predominant themes identified in the posts varied somewhat by patient-type (Figure 3): Patients were predominantly focused on their disease (including site), diagnostic- and treatment-related issues; caregivers were focused on side effects, which family member was diagnosed, and a treatment plan; and the majority of sentiment for Others related to sympathy and well-wishes.
- The performance of BERT vs. traditional models is terms of precision, recall, and F1 score is set forth in **Table 1**. The BERT approach had the highest precision, recall, and F1 score in all but one instance in which the naïve Bayes had greater precision for the Other subgroup.

Objectives

- To evaluate the ability of the BERT large language model (LLM) to classify health-related social media posts identified within online NSCLC patient forums by person-type (i.e., patients, caregivers, others).
- To compare performance of an optimized BERT LLM with traditional text classification methods in terms of the ability to identify distinct linguistic patterns across person-types.

Methods

DATA

COLLECTION

Figure 1. Overview of the Study Methodology

Relevant sites for data collection were identified by searching for patient forums specifically related to NSCLC. Criteria for selection included site relevance to NSCLC discussions, active user participation, and the presence of posts from both patients and caregivers sharing their experiences.



From the collected dataset, a subset of 2,400 posts (800 posts from each of the three person-types: patients, caregivers, and others) were manually annotated, and used to train the model and capture the distinct linguistic patterns associated with each person-type.



BERT was fine-tuned to classify posts by person-type (patients, caregivers, others), capturing subtle language differences. Naive Bayes, Random Forest, and XGBoost were also trained using TF-IDF and bag-of-words for comparison. The same test set was used for BERT and traditional models, with an 80/20 train-test data split.



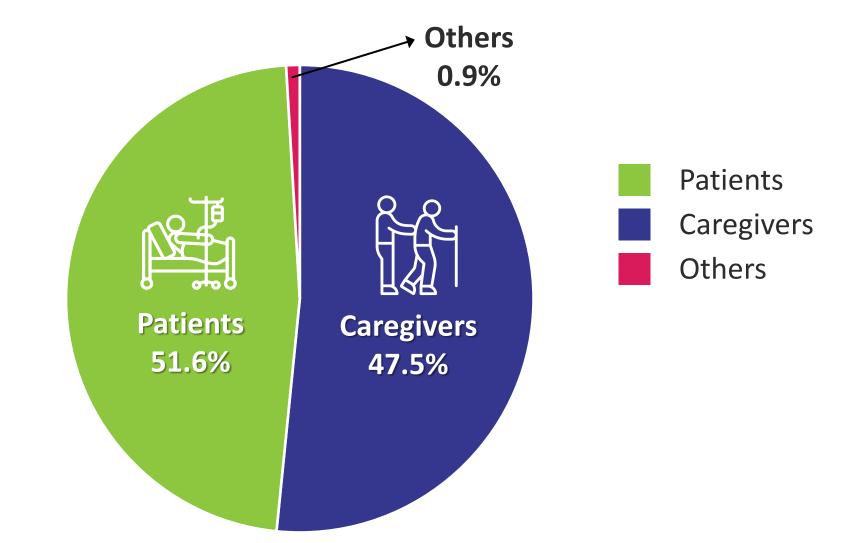
BERT performance was compared with traditional methods in terms of precision*, recall**, and F1 score***. These evaluation metrics collectively identify the ability of each model to consider subtle variations in language present in the data as it allocated each post to a given person-type.

Table 1. Performance Metrics of Machine Learning Models for Post Classification

Model Name	Precision (%) (Patient)	Recall (%) (Patient)	F1 Score (%) (Patient)	Precision (%) (Caregiver)	Recall (%) (Caregiver)	F1 Score (%) (Caregiver)	Precision (%) (Other)	Recall (%) (Other)	F1 score (%) (Other)	Total F1 (%) Score
Naive Bayes	82	86	84	62	98	76	98	36	52	71
Random Forest	84	66	74	83	72	77	61	82	70	73
XGboost	86	82	84	87	85	86	76	81	78	83
BERT	94	93	93	91	95	93	89	86	88	91

Abbreviation: BERT = Bidirectional encoder representations from transformers

Figure 3. Distribution of User Voices After Categorization



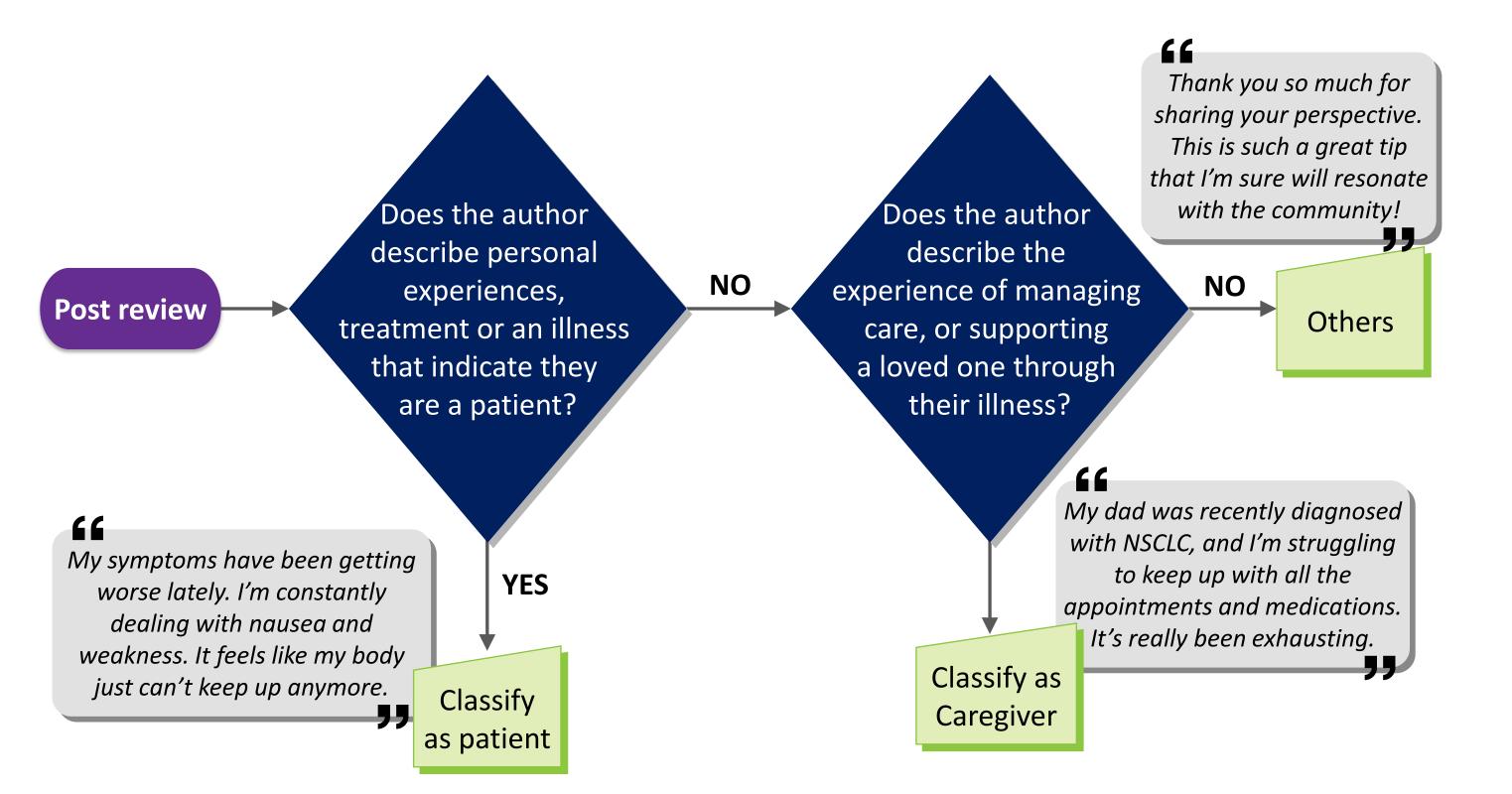


After training, the fine-tuned BERT model was deployed to perform inference on the remaining posts. The model analyzed these posts and classified them by user-type (patients, caregivers, others), utilizing the learned linguistic patterns to infer the appropriate patient-type category for each post.

* Precision: How many of the predicted positive results were actually correct.**Recall: How many of the actual positive cases were correctly identified.*** F1 Score: The balance between precision and recall.

Abbreviations: BERT = Bidirectional encoder representations from transformers; NSCLC = non-small cell lung cancer; TF-IDF = term frequency-inverse document frequency





The percentages represent the proportion of users classified into each category. The total number of users in each class is also provided for reference. Patients: 51.6% (N = 1,102 users), Caregivers: 47.7% (N = 1,014 users) and Others: 0.9% (N = 19 users).

Figure 4. Word Cloud of Themes from Patients, Caregivers, and Others

stav strond lagnosed stage sorry hea cancer spread Pairs ago Weeks ago days lab et KNOW Hope goes Ass Odd OOOD Ukke time **next** week isgnosed Scan Stage luck two weeks ing time terms lafficul time left lung ence disease brain tumor view growing encer SIC Blood work ETTECTS weeks clinical trial sounds like make sure long time brain mets nce side effects nusband diagnosed. Ivmph nodes feel better diagnosed lung mum diagnosed lung stage lung right lung ad Oont Know ts love treatment plan day time accordigist Stay course luna PATIENTS **CAREGIVERS Discussion and Conclusion** BERT outperformed traditional models in the ability to classify patient and caregiver voices as identified in social media posts, proving its value in analyses of these data. Most users were patients or caregivers, indicating that social media forums attract key stakeholders directly impacted by the disease.

Conversations across all groups were varied but tended to focus on selected issues that varied by patient-type. Patients discussed treatments, appointments, and illness; caregivers focused on family members; and those in the "Other" category primarily offered encouragement and support, reflecting broader community involvement.

Abbreviation: NSCLC = non-small cell lung cancer

User-Level Categorization Process

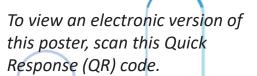
- Voice/Post Class Prediction: Posts were classified as "Patient," "Caregiver/Family," or "Other." For users whose posts were classified into multiple voices, a single predominant voice was assigned to reflect their overall role.
- **Counting Post Occurrences:** Posts were grouped by user identification (ID) to count how often each user posted in each person-type category.
- **Deprioritizing "Other" Class:** If a user had both "Other" and "Patient" or "Caregiver/Family" posts, the "Other" category was ignored. Only users with exclusively "Other" posts were classified as "Other."
- Assigning the Majority Class: Users were assigned the category they posted in most frequently. If "Patient" and "Caregiver/Family" posts were present, the higher-frequency category was selected.
- While these findings are encouraging for classification of social media posts, further analyses, such as topic modelling and clinical review, are essential to uncover deeper themes from patient and caregiver social media data that can be converted to RWE that ultimately can improve the lives of those with NSCLC.

References

- 1. Tan WW, Huq S. Non-Small Cell Lung Cancer. Medscape. Updated June 22, 2023. Accessed October 7, 2024. https://emedicine.medscape.com/article/279960-overview.
- 2. Cordoş AA, Bolboacă SD, Drugan C. Social media usage for patients and healthcare consumers: A literature review. Publications. Publications. 2017;5(2):9.
- 3. Real Life Sciences. 2020. FDA regulations and social media patient experience research. Real Life Sciences. https://rlsciences.com/fda-regulations-and-social-media-patient-experience-research/

Acknowledgment

Editorial and graphic design support were provided by Michael Grossi and Kawthar Nakayima of Evidera, a bus ness unit of PD, part of Thermo Fisher Scientific



hrough QR code are for

be reproduced without

Funding provided by Evidera Inc, a business unit of PPD, part of Thermo Fisher Scientific



Best wishes

et know

OTHERS

Presented at the ISPOR Europe Conference • 17–20 November 2024 • Barcelona, Spain