

Evidera PPD
Part of Thermo Fisher Scientific
MSR50

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

- 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.

- 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.

Figure 1. Overview of the Study Methodology

* 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

```

graph LR
    Start([Post review]) --> D1{Does the author describe personal experiences, treatment or an illness that indicate they are a patient?}
    D1 -- YES --> C1[Classify as patient]
    D1 -- NO --> D2{Does the author describe the experience of managing care, or supporting a loved one through their illness?}
    D2 -- YES --> C2[Classify as Caregiver]
    D2 -- NO --> C3[Others]
  
```

Post review

Does the author describe personal experiences, treatment or an illness that indicate they are a patient?

YES

Classify as patient

“My symptoms have been getting worse lately. I’m constantly dealing with nausea and weakness. It feels like my body just can’t keep up anymore.”

NO

Does the author describe the experience of managing care, or supporting a loved one through their illness?

YES

Classify as Caregiver

“My dad was recently diagnosed with NSCLC, and I’m struggling to keep up with all the appointments and medications. It’s really been exhausting.”

NO

Others

“Thank you so much for sharing your perspective. This is such a great tip that I’m sure will resonate with the community!”

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.

- 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 in 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.

Model Name	Precision (%)	Recall (%)	F1 Score (%)	Precision (%)	Recall (%)	F1 Score (%)	Precision (%)	Recall (%)	F1 score (%)	Total F1 (%)
	(Patient)	(Patient)	(Patient)	(Caregiver)	(Caregiver)	(Caregiver)	(Other)	(Other)	(Other)	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

A pie chart illustrating the distribution of users across three categories: Patients, Caregivers, and Others. The chart is divided into three segments: a large green segment for Patients (51.6%), a large dark blue segment for Caregivers (47.5%), and a very small pink segment for Others (0.9%). Each segment contains an icon: a hospital bed for Patients, two people for Caregivers, and a small pink square for Others. A legend to the right of the chart identifies the colors: green for Patients, dark blue for Caregivers, and pink for Others.

Category	Percentage	User Count (N)
Patients	51.6%	1,102
Caregivers	47.5%	1,014
Others	0.9%	19

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).

[illegible]

- 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.
- 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.

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*. 2017;5(2):9.
3. Real Life Sciences. 2020. FDA regulations and social media patient experience research. <https://rlsciences.com/fda-regulations-and-social-media-patient-experience-research/>

Editorial and graphic design support were provided by Michael Grossi and Kawthar Nakayima of Evidera, a business unit of PPD, part of Thermo Fisher Scientific.

To view an electronic version of this poster, scan this Quick Response (QR) code.

Copies of this poster obtained through QR code are for personal use only and may not be reproduced without permission from the congress and the author of this poster.

