**Rethinking Large Language Models Use Multivariate Statistics: Accuracy in Analysis of Sentiments of Physician Reviews**

Farrokh Alemi, Ph.D., others?

[falemi@gmu.edu](mailto:falemi@gmu.edu) 703 8933799

**Abstract**

**Background:** Large Language Models (LLMs) have shown high accuracy in sentiment analysis but training these systems to reduce rare errors have proven difficult, in part because of a large number of new concepts such as cosine similarity, positional probabilities, attention, transformer architecture, and neural networks that are not transparent about how they arrive at conclusions. **Objective:** We redo sentiment analysis using LLMs using simple and transparent statistical tools. The paper reports the accuracy of the two methods of arriving at sentiments of text. **Methods**: The statistical approach predicts the sentiment of target text using weighted regression models (where weights are based on similarity of phrases in target and training text). The paper reports the calibration of the statistical and LLM models in randomly selected 100 complaints and 100 praises. **Results**: The LLM and the similarity weighted regression analysis did not have statistically significant difference in calibration. The similarity weighted regression method was transparent and able to learn from an error by adding a single prompt to the training set. The LLM method was not transparent and did not provide a clear re-training strategy except to redo human ratings of a large number of cases. **Discussion**: Generative language models may need to rely on simpler similarity-based inference mechanism so that these models can be trained more efficiently.

**Introduction**

Generative large language models (LLMs) have been criticized for fabrication (errors in advice) and topic drift and investigators have emphasized the importance of training these models before using them. The training of these models depends on the inference engine used. This paper contrasts the ability to train two different inference engines. The LLM model is not transparent and it is not clear how it can be trained. The statistical model is transparent and a single case added to the training set corrects the error.

LLMs can be applied to prediction of a variety of outcomes, including generation of next word. We focus here on use of these models for sentiment analysis – a simple text processing task that has a single outcome and therefore it should be easy to compare the traditional and transparent statistical models to LLMs. There is widespread use of language models in analyzing sentiment of text [[[1]](#endnote-1)].

Human’s use of language is varied and innovative. Humans create many new combinations of words, not present in the training data sets, even when these sets are massive. It is not always possible to find an exact match to a target text within the training set. LLMs are able to transfer what is learned in training sets to a new target text, as seen in few-shot [[[2]](#endnote-2)], or zero-shot [[[3]](#endnote-3)] models. Generative Pre-trained Transformer architectures (e.g., GPT-4) have demonstrated highly accurate models for sentiment analysis. From time to time, errors occur [[[4]](#endnote-4),[[5]](#endnote-5)] and models must be retrained to avoid known errors.

Increasing the size of the training set, gathering more examples, improves accuracy but a plateau is reached beyond which accuracy may not improve, even when more data are collected [[[6]](#endnote-6)]. Since no method is perfect, to reduce errors, investigators have proposed to teach LLMs to learn from their errors. A key issue in selection of methods for sentiment analysis is how quickly these methods can learn from its errors.

Key feature of language models is reliance on neural networks, transformers, cosine similarity, attention, and concepts that have no support in probability theory. These methods look at possible combinations of words and make inferences that are highly accurate but it is not clear how the inference is arrived at. Statistical methods include naïve Bayes [[[7]](#endnote-7),[[8]](#endnote-8)], logistic regression [[[9]](#endnote-9)], k-Nearest Neighbors [[[10]](#endnote-10),[[11]](#endnote-11)], Support Vector Machine [[[12]](#endnote-12),[[13]](#endnote-13),[[14]](#endnote-14),[[15]](#endnote-15)], and neural network models [[[16]](#endnote-16)], to name a few. In the statistical models, words in the training prompts are used to predict the classification (completion) of a target text. If then rules are a third method for sentiment analysis [[[17]](#endnote-17),[[18]](#endnote-18)], these methods find the most similar examples in the training prompts and uses the average sentiment of these prompts. This paper provides evidence on how these three methods compare in accuracy and in learning from their erros.

In healthcare, sentiment analysis arises in many settings, including: (1) patient reviews of clinicians [[[19]](#endnote-19),[[20]](#endnote-20)], (2) risk of suicide in text-based social media [[[21]](#endnote-21),[[22]](#endnote-22)], (3) patient satisfaction rating [[[23]](#endnote-23)], (4) assessment of health status from clinical narratives [[[24]](#endnote-24)] and (5) many other circumstances. Here we focus on analysis of patient reviews of physicians.

**Methods**

**Study Design:** This is a retrospective cohort design. In this paper, we refer to the new text that needs to be classified as target text/comment. Analytical methods need to transfer what is known in the training set to predicting the sentiment of these novel and new combination of words. An error is found when the predicted sentiment and human classification of the sentiment of the text do not agree.

**Source and Size of Data**: Since our focus is on speed of retraining from known errors, we sample data in a manner to identify more errors. Rapid Improvement Inc. collected patient comments about clinicians from its participating clinics. A human reviewer classified 122,081 comments about clinicians into complaint, praise, or neither. We consider these human classifications of texts as the gold standard, against which we can examine the performance of different methods. A random sample of 100 complaints and 100 praises were used as test set and the remaining comments were used as the training set. Since the vast majority of comments are positive, this approach over samples complaints.

Some comments in the training set are duplicates. When a comment was chosen to be included in the training set, then 1 occasion of this comment was dropped from the training set. This allowed for relatively rare situation that some comments in the training set were exact match to the comment in the test set.

Long comments were split into smaller phrases to improve accuracy of the classification.

Only the relevant portion of the training data were used to classify the target text. Approximately 4 out of 5 comments in the training set did not share a word with a selected target comment. Obviously, only training cases that share a word with target comment are relevant. Including all comments would simply over state accuracy of methods. By focusing only on relevant sample, statistical models have a computationally easier ability to predict the sentiment of the comment.

In the regression method, separate regressions were constructed for each target comment. In these regressions, the independent variables are words in the target comment. Note that these are the only words that can affect the classification of the target comment. The observations are relevant comments in the training set. The response variable is the classification/completion of relevant comments. In the following equation, the index t refers to target comment t, and are parameters estimated from relevant portion of training data set. indicate words in the target comment and occasionally within the relevant training comments:

Since all the words are present in the target comment, the sentiment can be predicted as: . Each regression is used only once to predict one target comment, as the regression coefficients depend on different relevant portion of the training data. A repeated regression is necessary for analysis of the next target comment. This approach allows the words in the regression model to play different roles for different target comments. The target comment “The surgeon is knife happy” and “I am happy with this surgery” will have two different relevant data sets, two different regressions, allowing the model to fit and predict from only the relevant data. Also note that the inclusion of interaction terms allows combination of words to play a different role than individual words. In this repeated regression approach, the sequence of words plays no role in the classification task, thus “not happy to visit” and “happy not to visit” are considered the same.

In the second approach, a similarity score is used to find the example prompts in the training set that are most similar to the target comment. The average classification of the most similar comments in the training data is the classification for the target comment. Investigators modeling natural language often rely on cosine similarity to measure semantic similarity across two snippets of text [[[25]](#endnote-25)]. We prefer to use a similarity model shown to duplicate how human’s judgement of similarity [[[26]](#endnote-26),[[27]](#endnote-27)]:

is number of phrases matched in both target and training comments. A phrase is properly matched if it is in both comments and in right sequence. LLMs also require words to be in right positions.

is number of phrases in the target but not properly matched in the training comment.

is number of phrases not properly matched in target but present in the training comment

0 is a hyper parameter that shows the relative importance of missing phrases in target or training comments.

To understand these procedures, see Appendix A, where we present a Python code that identifies all phrases within a target or training comment. For the target comment “The surgeon is knife happy” the following phrases were identified after dropping the small word “The”:

1. Surgeon
2. Surgeon is
3. Surgeon is knife
4. Surgeon is knife happy
5. Is
6. Is knife
7. Is knife happy
8. Knife
9. Knife happy
10. happy

For the training comment “I am happy with this surgery”, the following phrases were identified:

1. I
2. I am
3. I am happy
4. I am happy with
5. I am happy with this
6. I am happy with this surgery
7. Am
8. Am happy
9. Am happy with
10. Am happy with this
11. Am happy with this surgery
12. Happy
13. Happy with
14. Happy with this
15. Happy with this surgery
16. With
17. With this
18. With this surgery
19. This
20. This surgery
21. Surgery

Note that the target comment “The surgeon is knife happy” share the word “happy” and two concepts “surgeon” matches with “surgery”, and “am” matches with “is.” Thus, . Many phrases such as “Knife” are in the target comment and not in the training comment; we calculated this to be There are also many phrases in the training comment that are not in the target comment, yielding . At alpha of 0.8, the similarity of these two texts is:

Note that when the target and training comments are exact matches, then the similarity score is 1.0, the highest possible value. When a short target is matched word by word to a long training comment, the similarity score will be a number less than 1, the longer the training comment the lower the score. Thus, additional unmatched word in the training comment lowers the similarity score. Also note that the similarity score is lowered when words in the target are not in the training set.

Because performance of language models is relatively high, it is important to find a measure of accuracy that is sensitive to small changes. There are many ways to examine the performance of models [[[28]](#endnote-28)]: Area under the Receiver Operating Characteristic curves, calibration using Brier scores [[[29]](#endnote-29)], reclassification tables [[[30]](#endnote-30)], and net reclassification improvement [[[31]](#endnote-31)]. We chose to use calibration because it is the least likely to be overwhelmed by imbalanced samples of data and most likely to report small differences in prediction probabilities. LLMs change in their predictions of the sentiment and we take the average of the first two estimates from LLMs.

**Results**

Table 1 shows the text of target comments, the size of relevant data for each target comment, and the result of predicting the sentiment from similar examples in the training set versus predicting it from regression models. If we look at area under the curve, there is always a cutoff that will classify 100% of comments in our test sample. The area under the curve for all four models will be near perfect predictions. If we examine calibration, relatively small differences among the model emerges. The average calibration of prompt-based architecture was 0.25 (standard deviation of 0.20) and the average calibration of linear regression was 0.24 (standard deviation of 0.26). Linear regression with interaction terms had the best average calibration at 0.20 (standard deviation of 0.16). None of the differences in calibration were statistically significant.

**Table 1: Predictions in a Sample of Complaints and Praises**

(Errors in prediction at cutoff of 0.50 are colored gray)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comment (Class)** | **Relevant Training Sample Size** | **Average of Most Similar Prompts** | **LLM Prediction of Sentiment** | **Similarity Weighted Regression** |
| Ridiculous (Complaint) | 80 | 1.00 |  | 0.90 |
| Definitely do not recommend. (Complaint) | 17,260 | 0.69 |  | 0.73 |
| I was trying to make an appointment with dr XXX got put on hold 3 times with no call back (Complaint) | 34,499 | 0.94 |  | 0.91 |
| This is discrimination. (Complaint) | 8 | 1.00 |  | 0.90 |
| Inexcusable this place is a complete scam. (Complaint) | 2,213 | 0.66 |  | 0.68 |
| They still have my money. (Complaint) | 1,964 | 0.75 |  | 0.77 |
| Do not go here. (Complaint) | 16,045 | 0.34 |  | 0.66 |
| and it is now simply crooked. (Complaint) | 2,321 | 0.24 |  | 0.60 |
| but I wish I didn't get the surgery at all. (Complaint) | 21,725 | 0.57 |  | 0.84 |
| I feel he should have asked my permission before bringing someone else into the room that is not a part of his staff. (Complaint) | 27,262 | 0.67 |  | 0.68 |
| I just think they could've had better procedures in place when dealing with those who had active panic attacks on the table. (Complaint) | 9,153 | 0.43 |  | 0.54 |
| For that reason, I'm giving this review 2 stars. (Complaint) | 1,749 | 0.69 |  | 0.44 |
| and the manipulating and lying that is going on in his office. (Complaint) | 4,773 | 0.91 |  | 0.75 |
| and replace it however, Dr XXXX refused to except any calls that evening (Complaint) | 22,627 | 0.44 |  | 0.63 |
| Further, no one seems to be forthcoming with any information concerning my case. (Complaint) | 10,766 | 0.53 |  | 0.60 |
| the response was to tell me to just wait in the exam room they had placed me in. (Complaint) | 6,445 | 0.53 |  | 0.57 |
| I will never ever go back to the Tampa office nor would I recommend them to friends or family! (Complaint) | 17,916 | 0.47 |  | 0.49 |
| And when I told her I felt intimidated and this causes me to have anxiety attacks that I have to take medication (Complaint) | 4,797 | 0.67 |  | 0.59 |
| She had called 911 and lied on me instead of just asking me to leave and then She had me escorted out (Complaint) | 7,554 | 0.80 |  | 0.80 |
| However, I am not happy with my rhinoplasty revision results. (Complaint) | 21,027 | 0.59 |  | 0.68 |
| Horrific personality, zero patient care, totally ignored me, and my medical condition impeding vision issues. (Complaint) | 14,836 | 0.48 |  | 0.64 |
| Unfortunately for me, that is Part of the procedure that I paid (Complaint) | 4,387 | 0.58 |  | 0.86 |
| However, I believe that my procedure was rushed because I don't think Dr. XXXX wanted to be there. (Complaint) | 37,427 | 0.52 |  | 0.71 |
| I asked for a prescription of regular hormones, she refused, I waited 2 weeks (Complaint) | 5,914 | 0.81 |  | 0.75 |
| And I won't go back. (Complaint) | 17.270 | 0.38 |  | 0.84 |
| Then your left hanging high and dry. (Complaint) | 2,254 | 0.72 |  | 0.60 |
| She did not care about any of that. (Complaint) | 22,136 | 0.78 |  | 0.91 |
| It was creepy and disgusting how he behaved and felt enabled. (Complaint) | 75 | 0.91 |  | 0.86 |
| Such a joke. (Complaint) | 69 | 1.00 |  | 0.62 |
| I'm still sick. (Complaint) | 1,326 | 0.63 |  | 0.72 |
| And I'm pretty sure any good reviews are flat out FAKE or patients who haven't had him cut them open (Complaint) | 25,073 | 0.25 |  | 0.75 |
| He never even checked on his patient. (Complaint) | 10,120 | 0.64 |  | 0.47 |
| Hardly any information was disseminated to me concerning glaucoma and its ramifications. (Complaint) | 3,140 | 0.41 |  | 0.50 |
| This leads me to believe the doctor really does not care or is completely blind to what she is doing (Complaint) | 37,471 | 0.80 |  | 0.58 |
| Shame on the whole ER unit that evening. (Complaint) | 1,105 | 0.55 |  | 0.64 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dr XXXX was great. (Praise) | 22,310 | 0.14 |  | 0.01 |
| I can't say enough great things about him his work (Praise) | 24,826 | 0.24 |  | 0.09 |
| I highly recommend him to anyone who is considering this procedure. (Praise) | 7,006 | 0.11 |  | 0.03 |
| and his office they made the entire process so smooth for me and seamless. (Praise) | 6,008 | 0.47 |  | 0.07 |
| I had good experience with Dr XXXX and his team (Praise) | 24,617 | 0.10 |  | 0.04 |
| I didn't feel any pain , it was a great experience (Praise) | 25,467 | 0.07 |  | 0.09 |
| for the first time as well as our 10-week appointment where we saw the baby dance for us! (Praise) |  | 0.33 |  | 0.11 |
| Very happy I found Dr. XXXX he did an awesome job on my hair transplant! (Praise) |  | 0.03 |  | 0.06 |
| He kept my husband and I well informed at all times and his staff was very friendly. (Praise) | 24,826 | 0.11 |  | 0.03 |
| I cannot thank everyone enough at XXXX for the help they've provided us. (Praise) |  | 0.25 |  | 0.09 |
| Loved my results. (Praise) |  | 0.06 |  | 0.02 |
| Doctor’s availability to talk and FaceTime before and after surgery was reassuring and calming. (Praise) |  | 0.16 |  | 0.11 |
| Great experience overall! (Praise) |  | 0.05 |  | 0.02 |
| and the clinic was able to see us same day to ensure baby was safe. (Praise) |  | 0.09 |  | 0.10 |
| Very good exam (Praise) |  | 0.36 |  | 0.13 |
| had both of my knee replaced and I am doing great. (Praise) |  | 0.06 |  | 0.13 |
| Dr. XXXX is excellent. (Praise) |  | 0.31 |  | 0.03 |
| highly recommended. (Praise) |  | 0.19 |  | 0.01 |
| XXXX spent time going over my labs and answered all of my questions. (Praise) |  | 0.11 |  | 0.20 |
| Friendly, courteous staff. (Praise) |  | 0.02 |  | 0.02 |
| She is an incredible injector! (Praise) |  | 0.18 |  | 0.18 |
| What would I do without XXXX! (Praise) |  | 0.04 |  | 0.24 |
| Happy (Praise) |  | 0.00 |  | 0.01 |
| So gentle, I barely feel a thing! (Praise) |  | 0.08 |  | 0.10 |
| XXXX is professional and so good at what she does! (Praise) |  | 0.06 |  | 0.06 |
| All staff and nurses are very helpful. (Praise) |  | 0.29 |  | 0.03 |
| Very streamlined process from start to finish. (Praise) |  | 0.20 |  | 0.19 |
| Doctor XXXX is clearly very skilled and knowledgeable, and I am very happy with my results! (Praise) |  | 0.03 |  | 0.03 |
| I appreciated that very much. (Praise) |  | 0.00 |  | 0.05 |
| Dr. XXXX was very honest with what I should expect. (Praise) |  | 0.07 |  | 0.06 |
| **Average Calibration Score** |  | **0.25** | **0.24** | **0.20** |
| **Standard Deviation of Calibration Scores** |  | **0.20** | **0.26** | **0.16** |

Table 2 provides data on how rapidly the various methods can be retrained to avoid their errors.

**[More to come]**

**Discussion**

**Needs Revision**

Repeated regression models in the relevant portion of data had a relatively high accuracy rate. At cutoff of 0.5, only one comment was misclassified. This shows that it is possible to organize statistical models that have relatively low error rates. These results are not fundamentally different from the literature and was expected. Many studies reported to date stop at reporting accuracy and fail to report how the model can be retrained to avoid the error in the future. In this paper, we examined the retraining effort.

Statistically-optimal, repeated, regression models had the same level of calibration as similarity scores. Therefore, these methods were not different in accuracy, despite differential levels of errors at different cutoffs. Yet, these models differed in other ways. Computationally, similarity-based inference was relying on scores that are relatively easier to calculate than regression coefficients. In addition, similarity-based architecture was able to learn from its errors in one shot, only one prompt-completion was needed to correct the model. In contrast, statistical models needed a large number of corrections. [More]

**References**

1. Alexander G, Bahja M, Butt GF. Automating Large-scale Health Care Service Feedback Analysis: Sentiment Analysis and Topic Modeling Study. JMIR Med Inform. 2022 Apr 11;10(4):e29385. doi: 10.2196/29385. PMID: 35404254; PMCID: PMC9039814. [↑](#endnote-ref-1)
2. Brown, Mann, Ryder. Language models are few-shot learners. Adv Neural Inf Process Syst n.d. [↑](#endnote-ref-2)
3. Socher R, Ganjoo M, Sridhar H, Bastani O, Manning CD, Ng AY. Zero-shot learning through cross-modal transfer. ArXiv [CsCV] 2013. [↑](#endnote-ref-3)
4. Peng C, Yang X, Yu Z, Bian J, Hogan WR, Wu Y. Clinical concept and relation extraction using prompt-based machine reading comprehension. ArXiv [CsCL] 2023. [↑](#endnote-ref-4)
5. Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A Unified MRC Framework for Named Entity Recognition. Presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. [↑](#endnote-ref-5)
6. Pacurari AC, Bhattarai S, Muhammad A, Avram C, Mederle AO, Rosca O, Bratosin F, Bogdan I, Fericean RM, Biris M, Olaru F, Dumitru C, Tapalaga G, Mavrea A. Diagnostic Accuracy of Machine Learning AI Architectures in Detection and Classification of Lung Cancer: A Systematic Review. Diagnostics (Basel). 2023 Jun 22;13(13):2145. doi: 10.3390/diagnostics13132145. PMID: 37443539; PMCID: PMC10340581. [↑](#endnote-ref-6)
7. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning. J Med Internet Res. 2015 Aug 25;17(8):e208. doi: 10.2196/jmir.4392. PMID: 26307512; PMCID: PMC4642404. [↑](#endnote-ref-7)
8. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets. JMIR Public Health Surveill. 2016 Oct 24;2(2):e162. doi: 10.2196/publichealth.6327. PMID: 27777215; PMCID: PMC5099500. [↑](#endnote-ref-8)
9. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets. JMIR Public Health Surveill. 2016 Oct 24;2(2):e162. doi: 10.2196/publichealth.6327. PMID: 27777215; PMCID: PMC5099500. [↑](#endnote-ref-9)
10. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning. J Med Internet Res. 2015 Aug 25;17(8):e208. doi: 10.2196/jmir.4392. PMID: 26307512; PMCID: PMC4642404. [↑](#endnote-ref-10)
11. Myslín M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res. 2013 Aug 29;15(8):e174. doi: 10.2196/jmir.2534. PMID: 23989137; PMCID: PMC3758063. [↑](#endnote-ref-11)
12. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning. J Med Internet Res. 2015 Aug 25;17(8):e208. doi: 10.2196/jmir.4392. PMID: 26307512; PMCID: PMC4642404. [↑](#endnote-ref-12)
13. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets. JMIR Public Health Surveill. 2016 Oct 24;2(2):e162. doi: 10.2196/publichealth.6327. PMID: 27777215; PMCID: PMC5099500. [↑](#endnote-ref-13)
14. Myslín M, Zhu SH, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res. 2013 Aug 29;15(8):e174. doi: 10.2196/jmir.2534. PMID: 23989137; PMCID: PMC3758063. [↑](#endnote-ref-14)
15. Sofean M, Smith M. Sentiment analysis on smoking in social networks. Stud Health Technol Inform. 2013;192:1118. PMID: 23920892. [↑](#endnote-ref-15)
16. Seilsepour A, Ravanmehr R, Nassiri R. Topic sentiment analysis based on deep neural network using document embedding technique. J Supercomput. 2023 Jun 5:1-39. doi: 10.1007/s11227-023-05423-9. Epub ahead of print. PMID: 37359345; PMCID: PMC10241384. [↑](#endnote-ref-16)
17. Desai T, Shariff A, Shariff A, Kats M, Fang X, Christiano C, Ferris M. Tweeting the meeting: an in-depth analysis of Twitter activity at Kidney Week 2011. PLoS One. 2012;7(7):e40253. doi: 10.1371/journal.pone.0040253. Epub 2012 Jul 5. PMID: 22792254; PMCID: PMC3390326. [↑](#endnote-ref-17)
18. Tighe PJ, Goldsmith RC, Gravenstein M, Bernard HR, Fillingim RB. The painful tweet: text, sentiment, and community structure analyses of tweets pertaining to pain. J Med Internet Res. 2015 Apr 2;17(4):e84. doi: 10.2196/jmir.3769. PMID: 25843553; PMCID: PMC4400316. [↑](#endnote-ref-18)
19. Tang JE, Arvind V, White CA, Dominy C, Kim JS, Cho SK, Walsh A. Using Sentiment Analysis to Understand What Patients Are Saying About Hand Surgeons Online. Hand (N Y). 2021 Dec 30:15589447211060439. doi: 10.1177/15589447211060439. Epub ahead of print. PMID: 34969297. [↑](#endnote-ref-19)
20. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual Manag Health Care. 2012 Jan-Mar;21(1):9-19. doi: 10.1097/QMH.0b013e3182417fc4. PMID: 22207014. [↑](#endnote-ref-20)
21. Kadriu F, Murtezaj D, Gashi F, Ahmedi L, Kurti A, Kastrati Z. Human-annotated dataset for social media sentiment analysis for Albanian language. Data Brief. 2022 Jul 2;43:108436. doi: 10.1016/j.dib.2022.108436. PMID: 35832321; PMCID: PMC9272335. [↑](#endnote-ref-21)
22. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res. 2021 May 4;23(5):e15708. doi: 10.2196/15708. PMID: 33944788; PMCID: PMC8132982. [↑](#endnote-ref-22)
23. Hopper AM, Uriyo M. Using sentiment analysis to review patient satisfaction data located on the internet. J Health Organ Manag. 2015;29(2):221-33. doi: 10.1108/JHOM-12-2011-0129. PMID: 25800334. [↑](#endnote-ref-23)
24. Sanglerdsinlapachai N, Plangprasopchok A, Ho TB, Nantajeewarawat E. Improving sentiment analysis on clinical narratives by exploiting UMLS semantic types. Artif Intell Med. 2021 Mar;113:102033. doi: 10.1016/j.artmed.2021.102033. Epub 2021 Feb 12. PMID: 33685589. [↑](#endnote-ref-24)
25. Ormerod M, Martínez Del Rincón J, Devereux B. Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis. JMIR Med Inform. 2021 May 26;9(5):e23099. doi: 10.2196/23099. PMID: 34037527; PMCID: PMC8190645. [↑](#endnote-ref-25)
26. Rahnama J, Hüllermeier E. Learning Tversky Similarity. Information Processing and Management of Uncertainty in Knowledge-Based Systems. 2020 May 15;1238:269–80. doi: 10.1007/978-3-030-50143-3\_21. PMCID: PMC7274714. [↑](#endnote-ref-26)
27. Tversky, A. (1977). Features of similarity. Psychological Review, 84(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327 [↑](#endnote-ref-27)
28. Steyerberg EW, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21: 128-38. [↑](#endnote-ref-28)
29. Brier GW. "Verification of Forecasts Expressed in Terms of Probability" (PDF). Monthly Weather Review 1950. 78: 1–3. [↑](#endnote-ref-29)
30. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115: 928-35. [↑](#endnote-ref-30)
31. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27: 157-72. [↑](#endnote-ref-31)