Machine Learning Classification Using Emotion Language and Subjective Difficulty to Distinguish Patients with Functional Seizures from Trauma Controls

Natalie N. NEWTON[a,1], Nicole A. ROBERTS[a], Estrella M. CONTRERAS[a] and Mary H. BURLESON[a]
[a] *School of social and Behavioral Sciences, Arizona State University, Tempe, Arizona, USA.*
ORCiD ID: Natalie N. Newton *https://orcid.org/0000-0002-3905-4423*

Abstract. Functional seizures (FS) is a clinical condition where individuals experience seizure-like symptoms without the expected electrocortical basis for them. Those with FS tend to experience difficulty with negative emotions and have trouble labeling or describing their emotions. We used collected data from a relived emotion task in which FS and a trauma control group (TC) wrote descriptions when prompted to recall memories evoking anger and shame feelings and subsequently rated subjective difficulty for the reliving task. We conducted a sentiment analysis by matching words from descriptions to words in the Affective Norms for English Words (ANEW) database to compute language features of word count, valence, and arousal. Using language features and subjective task difficulty, we tested 45 machine learning models with a logistic regression classification engine to test which features worked best to distinguish FS from TC. For models with language features only, features of the relived shame condition were more accurate at distinguishing FS from TC; models with word count and with valence and/or arousal added were more accurate than with word count alone. Models with language features and difficulty were better at distinguishing FS from TC than models with language features alone. However, the two models with subjective difficulty features alone emerged as the most accurate to distinguish FS from TC. This serves as a valuable first step at demonstrating the utility of language features in combination with subjective ratings at discerning FS from TC, although more work is needed to bolster contributions of language features to this type of ML modeling.

Keywords. Functional seizures, sentiment analysis, machine learning, emotion

1. Introduction

Functional seizures (FS), or psychogenic non-epileptic seizures, is a clinical condition in which patients experience seizure-like symptoms without the expected aberrant electrocortical activity during episodes [1-3]. FS are diagnosed in under 1% of the general population, but FS patients make up ~20% of those seeking epilepsy treatment [4]. FS patients do not respond to anti-seizure medication and often have comorbid psychological conditions [5], making FS difficult to treat because the condition is primarily psychological [2, 3]. Nevertheless, FS is just as debilitating as epileptic seizures [6].

While the cause of FS is unclear, prior work suggests that FS patients tend to struggle with describing and processing emotions [7]. These difficulties, along with possible disruptions in interoception, agency, and greater "fight-or-flight" tendencies, may translate into abnormal motor movement [2, 8-10]. FS patients also tend to exhibit alexithymia (difficulty labeling/describing emotions), report more difficulty with negative emotions, and show more FS symptoms from exposure to negative stimuli [11-

[1] Corresponding Author: Natalie N. Newton, nnnewton@asu.edu

13]. Emotional processing issues and alexithymic tendencies make FS language use of particular interest.

A handful of studies have explored language and narratives produced by FS patients.

FS had more trouble recalling details during seizure episodes compared to epilepsy patients [14]. A qualitative study compared narratives produced by FS and epilepsy patients; FS tended to produce shorter, more negative narratives expressing powerlessness primarily stemming from their FS diagnosis [15, 16]. A smaller study found that FS patients commonly disclosed traumatic events but did not recognize them as such [17].

No study to our knowledge has leveraged natural language processing (NLP) to examine language produced by those with FS compared to a clinical control group on their past autobiographical or relived emotional memories. Language features can be used to differentiate normative populations from those with depressive symptoms [18] or those with posttraumatic stress disorder [19], demonstrating the utility of language among those with clinical conditions. Moreover, language use can shape emotional experiences [20, 21].

We investigated systematic differences in language use by FS compared to trauma controls (TC) when prompted to describe memories evoking anger and shame, respectively. We identified language features based on theoretical and empirical grounds to then use in supervised machine learning models as a means of classifying the two groups (FS vs. TC). We also added to the model's participant ratings of subjective (self-reported) difficulty in reliving negative emotions, as these demonstrated a strong group difference effect in our prior work [12]. We hypothesized that 1) models using language features and/or subjective ratings derived from the shame condition would be more accurate than models using only anger features. Additionally, we hypothesized that 2) language features, including word count and two defining characteristics of affective states, namely valence and arousal [25], would yield more accurate models to differentiate FS from TC than models with word count alone, and 3) models with subjective difficulty added as a feature would also yield greater accuracy compared to models with language features alone.

2. Methods

We examined language features and subjective difficulty ratings for a relived emotion task among FS and TC using supervised machine learning (ML) models. We used data collected from participants (N=60) with FS (N=11) or TC with varying levels of posttraumatic stress (PTS; N=49) from a larger study where they completed a relived emotion task [12, 22].

After a *neutral* condition, participants were asked to recall a time they felt strong amounts of *anger*, *shame*, or *happiness* (counterbalanced). For each condition, participants were instructed to write descriptions (~4 sentences) of the memory then verbalize it and rate the subjective difficulty of reliving the memory. Here we focused on written descriptions evoked from *anger* and *shame* conditions and the accompanying subjective difficulty.

2.1. Text and data analysis

We used an NLP technique to identify individual words, irrespective of order, used in written descriptions. Using the Tidytext package in R [23], descriptions were preprocessed by lemmatizing (e.g., *run*, *ran*, *running* all become *run*), removing punctuation and numbers, converting all words to lowercase, and extracting individual words used as tokens (the smallest unit of analysis). We did not account for polysemic words with multiple meanings, negation (e.g., "today was *not* great"), misspelled words, related emotion words (e.g., *monotony* vs. *monotonous*), or initialisms (e.g., *DUI*).

We conducted a sentiment analysis by matching words from descriptions to words found in the Affective Norms for English Words (ANEW) database, comprising 13,000+ words with crowdsourced valence and arousal ratings [24]. Emotions can be classified by valence (how positive or negative the emotion is) and arousal (how physiologically arousing or calming it is) dimensions [25]; single words can also be classified this way [24].

Using only description words found in ANEW, we computed word count, mean valence, and mean arousal per participant per emotion condition of *anger* and *shame*. Using 2 (emotion condition: *shame*, *anger*) x 2 (group: FS, TC) mixed measures ANOVA in SPSS with emotion condition as a repeated measure and group as a between-subject factor, we examined differences in word count, valence, arousal, and difficulty. Of note, 3 FS participants who did not complete the relived emotion task for either *anger* or *shame* did not provide difficulty ratings, and thus were excluded from comparisons of difficulty ratings and from subsequent ML analyses. For the machine learning (ML) classification models, we split data into training and test sets using a proportion split of .65 for the training data [26]. Because the proportion of FS vs. TC was imbalanced, we used stratification to ensure an even proportion of FS was in the training and test data.

Using supervised ML with a binary logistic regression classification engine with 5-fold cross-validation in R with the tidyModels package [27], we trained models and evaluated performance on test data using the area under the curve (AUC) and balanced accuracy (*BA*) values (calculated with the R caret package) [28, 29]. We tested 45 models with features of word count, valence, arousal, and difficulty ratings for *anger* and *shame* conditions individually and in combination (i.e., *anger* + *shame*).

3. Results

ANOVA results are in **Table 1**. Trend-level effects of emotional condition emerged for valence and word count. A significant group effect also emerged for valence and word count, with FS using fewer and more negative words than TC. For subjective ratings, there was a significant group effect, with FS reporting greater difficulty than TC. No other significant effects emerged.

3.1. Machine learning: Language features

Models that included only language features (word count, valence, and/or arousal) did not exceed a test data AUC of .80, with only some models achieving applicable test data *BA* values due to low sensitivity in identifying FS (see Table 2) [29]. Models with *BA*=.89 or greater emerged as those with the highest predictive accuracy at identifying FS. The model using word count and arousal for the *shame* condition (2 features) performed the best given its high *BA* and acceptable AUC (*BA*=.89, AUC=.72). The model with word count and arousal for the two conditions combined (*anger* + *shame*; 4 features) had higher sensitivity but poor AUC (*BA*=.92, AUC=.59). Although several models achieved an acceptable AUC of <.70, many of these models had poor sensitivity and low or not applicable *BA* [29, 30].

The other high performing models had equally high *BA* but lower AUC of <.70. The model with word count and valence for *anger* (2 features), valence and arousal for *shame* (2 features), and word count, valence, and arousal for *shame* (3 features) all performed the same (*BA*=.89, AUC=.68). Models with language features of word count, valence, or arousal in isolation (1 feature) for any of the emotion conditions were unable to identify FS participants. Since most higher performing models included features for *shame*, this in part supports our first hypothesis. Our second hypothesis also received support, given that valence and/or arousal language features improved model accuracy versus word count alone; however, neither valence nor arousal alone were adequate models.

Table 1. Comparisons of Valence, Arousal, Word Count, and Difficulty Ratings by Emotion Condition and Group

Outcome Measure	*df*	*F*-value	*p*-value	*ηp2*
Valence				
Emotion	1, 58	3.73	.058	.06
Group	1, 58	4.37	**.041**	.07
Emotion x Group	1, 58	2.12	.151	.04
Arousal				
Emotion	1, 58	0.80	.375	.01
Group	1, 58	0.13	.723	>.01
Emotion x Group	1, 58	0.82	.369	.01
Word Count				
Emotion	1, 58	2.89	.095	.05
Group	1, 58	11.67	**.001**	.17
Emotion x Group	1, 58	2.21	.143	.04
Difficulty Ratings				
Emotion	1,55	0.06	.801	.00
Group	1,55	15.52	**.001**	.22
Emotion x Group	1,55	0.79	.378	.01

Note: df = degrees of freedom; ηp^2 = effect size measured as *partial eta squared.*

3.2. Machine learning: Language features and subjective difficulty ratings

By adding rated difficulty of reliving emotional experiences to models with language features, predictive accuracy improved for most models, supporting our third hypothesis. Models with *BA*=.89 or greater and AUC³.70 emerged with the highest predictive accuracy at distinguishing FS from TC (**Table 2**). The model with difficulty ratings, word count, and arousal for *anger* (3 features) had the highest performance (*BA*=.92, AUC=.73). This was closely followed by a model with difficulty, valence, and arousal for *shame* (3 features; *BA*=.89, AUC=.76), as well as the model with difficulty and valence for *anger* (2 features; *BA*=.92, AUC=.70) and a model with difficulty, word count, valence, and arousal for *anger* (4 features; *BA*=.92, AUC=.70). Since most of the highest performing models did not include *shame* features, our first hypothesis was not supported when difficulty ratings were added to language features. Several other difficulty and language feature models fared well at identifying those with FS (*BA*=.89) but had lower AUC values (AUC<.70; see **Table 2**).

Using only subjective difficulty yielded the best performing models overall, outperforming those with language features. The model with difficulty for *anger + shame* (2 features) performed the best at distinguishing FS from TC (*BA*=.94, AUC=.84). The model with difficulty for *anger* (1 feature) performed similarly (*BA*=.94, AUC=.83).

Although the model with difficulty for *shame* had a comparable AUC value, its sensitivity at identifying FS was too low (*BA*=NA, AUC=.81). Thus, our first hypothesis (i.e., that *shame* models would outperform *anger* models) was not supported, and our third hypothesis was partially supported, insofar as subjective difficulty, albeit without language features, yielded the two best ML models. Given the small sample, models with 2 features or fewer are more likely to generalize to a larger sample than models with 4 features. The most accurate models generally overfit training data, which also warrants caution.

Table 2. Machine learning models results

	Emotion Condition								
	Anger			Shame			Anger + Shame		
Model features	**AUC Training**	**AUC Test**	**BA**	**AUC Training**	**AUC Test**	**BA**	**AUC Training**	**AUC Test**	**BA**
Word count	0.81	0.76	NA	0.92	0.76	NA	0.91	0.76	NA
Valence	0.73	0.52	NA	0.72	0.79	NA	0.75	0.68	NA
Arousal	0.79	0.39	NA	0.68	0.52	NA	0.93	0.52	NA
Word count, valence	0.85	0.68	0.89	0.93	0.79	NA	1.00	0.50	NA
Word count, arousal	0.88	0.57	NA	0.91	0.72	0.89	0.99	0.59	**0.92**
Valence, arousal	0.82	0.39	NA	0.68	0.68	0.89	0.93	0.47	0.55
Word count, valence,	0.88	0.61	NA	0.91	0.68	0.89	1.00	0.63	0.75
Difficulty	0.94	0.83	**0.94**	0.58	0.81	NA	0.94	0.84	**0.94**
Difficulty, word count	0.95	0.89	0.81	0.93	0.72	NA	0.98	0.84	0.37
Difficulty, valence	1.00	0.7	**0.92**	0.75	0.89	NA	1.00	0.77	0.75
Difficulty, arousal	0.98	0.63	**0.92**	0.72	0.63	0.89	1.00	0.63	0.89
Difficulty, word count,	1.00	0.68	**0.92**	0.73	0.81	NA	1.00	0.67	0.89
Difficulty, word count,	0.99	0.73	**0.92**	0.92	0.67	0.89	1.00	0.59	0.55
Difficulty, valence, arousal	1.00	0.60	0.89	0.73	0.76	0.89	1.00	0.73	0.75
Difficulty, word count,	1.00	0.70	**0.92**	0.91	0.64	NA	1.00	0.6	0.64

Note: AUC = Area under the curve; *BA* = test data balanced accuracy.

4. Discussion

FS have gained increased empirical attention, given the need to better understand, diagnose and treat this condition. The ML models presented here serve as preliminary evidence that language features in combination with subjective difficulty experiencing emotions can distinguish FS from a comparison group with prior trauma exposure and varying clinical symptom levels. Because those with FS tend to exhibit alexithymia [2, 8] and struggle with emotional processing difficulties for negative emotions [12], we expected language features, particularly emotion words reflecting valence and arousal, in addition to overall word count, to be consequential in ML models. Although language was not more relevant than subjective difficulty ratings in differentiating FS and TC, both sets of features combined – and even subjective task difficulty alone – were predictive

We examined features in the context of *anger* and *shame*, which are theoretically relevant to FS and to traumatic stress conditions. Since prior work has revealed that FS struggle with shame in particular [8, 12], we hypothesized that language features for *shame* would lead to more accurate ML models than those using only *anger* features. This was largely supported, since models with *shame* language features fared better than those using *anger* features. However, this changed when adding subjective difficulty to models with language features. Contrary to our first hypothesis, models using *shame* language and difficulty features were not among the top performing models when difficulty was added. In fact, the two models with subjective difficulty alone for *anger* and *anger + shame* were the best performing models tested, as they were able to distinguish FS from non-FS participants. This suggests that shame may be relevant in how those with FS communicate past events and emotions, whereas reported difficulty reliving anger, in addition to written expressions of anger-inducing experiences, may be relevant to characterizing FS versus TC.

Models with only 2 language features also performed well at identifying FS. While valence and arousal performed adequately combined and with word count, each performed poorly individually. Thus, awareness of both amount and content of text used may be relevant to clinicians and researchers aiming to differentiate those with and without FS. While several of the top performing models had acceptable or excellent AUC

values, all models should ideally have higher sensitivity in identifying FS [30]. High AUC along with high sensitivity and high *BA* are necessary [29].

The results should be interpreted with caution given several limitations. First, as noted earlier, polysemic words, proper nouns, initialisms, negation, misspelled words, and word discrepancies were not accounted for in text preprocessing. Possible cultural, individual stylistic, or other context-based differences in language use, including potential linguistic biases in how words are normed, also were not considered here. Topic content was also not considered but can bring up privacy concerns for patients discussing emotional memories with clinicians. Second, the sample was quite small and highly imbalanced, which makes our results including ML models with 6 features or more less likely to generalize to a larger sample. Third, we only examined language features and difficulty for the negative relived emotions; while these may be particularly challenging for those with FS, they yield salient, unpleasant memories that neither FS [12] nor those with PTS symptoms [31] are keen to experience, potentially minimizing group differences. Other language features may be relevant in differentiating groups, as might verbalizations beyond written text. Finally, other comparison groups may be relevant, such as those with epilepsy, particularly given that not everyone with FS has experienced trauma [2, 11].

With a larger dataset to train ML models and with other comparison groups, this work can be expanded and leveraged to create a training intervention and/or tool to help clinicians better identify whether a patient is experiencing FS (vs. epilepsy) and/or whether a patient is more avoidant of negative emotions. The current study demonstrated that subjective difficulty alone for negative relived emotions created the best ML models. When patients verbalize their emotional experiences, linguistic features may provide additional insight into their social and emotional interactions to distinguish FS from TC.

References

[1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. DSM-V. 2013. doi: https://doi.org/10.1176/appi.books.9780890425596

[2] Brown RJ, Reuber M. Towards an integrative theory of psychogenic non-epileptic seizures (PNES). *Clin Psychol Rev.* 2016;47:55-70. doi: https://doi.org/10.1016/j.cpr.2016.06.003

[3] LaFrance WC Jr, Baker GA, Duncan R, Goldstein LH, Reuber M. Minimum requirements for the diagnosis of psychogenic nonepileptic seizures: a staged approach: a report from the International League Against Epilepsy Nonepileptic Seizures Task Force. *Epilepsia.* 2013;54(11):2005-2018. doi: https://doi.org/10.1111/epi.12356

[4] Benbadis SR, Hauser WA. An estimate of the prevalence of psychogenic non-epileptic seizures. *Seizure.* 2000;9(4):280-281. doi: https://doi.org/10.1053/seiz.2000.0409

[5] Fiszman A, Alves-Leon SV, Nunes RG, Isabella DA, Figueira I. Traumatic events and posttraumatic stress disorder in patients with psychogenic nonepileptic seizures. *Epilepsy Behav.* 2004;5(6):818-825. doi: https://doi.org/10.1016/j.yebeh.2004.08.011

[6] Novakova B, Howlett S, Baker R, Reuber M. Emotion processing and psychogenic non-epileptic seizures: a cross-sectional comparison of patients and healthy controls. *Seizure.* 2015;29:4-10. doi: https://doi.org/10.1016/j.seizure.2015.03.005

[7] LaFrance WC Jr, Reuber M, Goldstein LH. Management of psychogenic nonepileptic seizures. *Epilepsia.* 2013;54(suppl 1):53-67. doi: https://doi.org/10.1111/epi.12106

[8] Drane DL, Fani N, Hallett M, Khalsa SS, Perez DL, Roberts NA. A framework for understanding the pathophysiology of functional neurological disorder. *CNS Spectr.* 2021;26(6):555-561. doi: https://doi.org/10.1017/S109285292000162X

[9] Roberts NA, Reuber M. Alterations of consciousness in psychogenic nonepileptic seizures: emotion, emotion regulation, and dissociation. *Epilepsy Behav.* 2014;30:43-49. doi:10.1016/j.yebeh.2013.09.040

[10] van der Kruijs SJ, Bodde NM, Vaessen MJ, Jansen JF. Functional connectivity of dissociation in patients with psychogenic non-epileptic seizures. *J Neurol Neurosurg Psychiatry.* 2012;83(3):239-247. doi: https://doi.org/10.1136/jnnp-2011-300776

[11] Kaplan MJ, Dwivedi AK, Bowman M. Comparisons of childhood trauma, alexithymia, and defensive styles in patients with psychogenic non-epileptic seizures. *J Psychosom Res.* 2013;75(2):142-146. doi: https://doi.org/10.1016/j.jpsychores.2013.05.007

[12] Roberts NA, Burleson MH, Torres DL, Wang NC. Emotional reactivity as a vulnerability for psychogenic nonepileptic seizures. *J Neuropsychiatry Clin Neurosci.* 2020;32(1):95-100. doi: https://doi.org/10.1176/appi.neuropsych.18070160

[13] Pick S, Millman LM, Ward E, David AS. Unravelling the influence of affective stimulation on functional neurological symptoms. *J Neurol Neurosurg Psychiatry.* 2024;95(5):461-470. doi: https://doi.org/10.1136/jnnp-2023-330123

[14] Cornaggia CM, Gugliotta SC, Magaudda A, Alfa R, Beghi M, Polita M. Conversation analysis in the differential diagnosis of Italian patients with epileptic or psychogenic non-epileptic seizures: a blind prospective study. *Epilepsy Behav.* 2012;25(4):598-604. doi: https://doi.org/10.1016/j.yebeh.2012.09.010

[15] Rawlings GH, Brown I, Reuber M. Narrative analysis of written accounts about living with epileptic or psychogenic nonepileptic seizures. *Seizure.* 2018;62:59-65. doi:10.1016/j.seizure.2018.09.017
[16] Rawlings GH, Brown I, Stone B, Reuber M. Written accounts of living with epilepsy or psychogenic nonepileptic seizures: a thematic comparison. *Qual Health Res.* 2018;28(6):950-962. doi: https://doi.org/10.1177/1049732318759930
[17] Peacock M, Dickson JM, Bissell P, Grunewald R, Reuber M. Beyond the medical encounter: can the free association narrative interview method extend psychosocial understandings of non-epileptic attack disorder? *J Psychosoc Stud.* 2022;15(1):36-51. doi: https://doi.org/10.1332/147867321X16119281118515
[18] Raffaelli Q, Mills C, de Stefano NA, O'Connor MF. The think aloud paradigm reveals differences in the content, dynamics, and conceptual scope. *Sci Rep.* 2021;11(1):19362. doi:10.1038/s41598-021-98876-3
[19] Sawalha J, Yousefnezhad M, Shah Z, Brown MR, Greenshaw AJ, Greiner R. Detecting presence of PTSD using sentiment analysis from text data. *Front Psychiatry.* 2022;12:811392. doi: https://doi.org/10.3389/fpsyt.2021.811392
[20] Lindquist KA, MacCormack JK, Shablack H. The role of language in emotion: predictions from psychological constructionism. *Front Psychol.* 2015;6:444. doi: https://doi.org/10.3389/fpsyg.2015.00444
[21] Barrett LF, Lindquist KA, Gendron M. Language as context for the perception of emotion. *Trends Cogn Sci.* 2007;11(8):327-332. doi: https://doi.org/10.1016/j.tics.2007.06.003
[22] Levenson RW, Carstensen LL, Friesen WV, Ekman P. Emotion, physiology, and expression in old age. *Psychol Aging.* 1991;6(1):28-35.
[23] Silge J, Robinson D. tidytext: text mining and analysis using tidy principles. *J Open Source Softw.* 2016;1(3):37. doi: https://doi.org/10.21105/joss.00037
[24] Warriner AB, Kuperman V, Brysbaert M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res Methods.* 2013;45(4):1191-1207. doi: https://doi.org/10.3758/s13428-012-0314-x
[25] Russell JA. A circumplex model of affect. *J Pers Soc Psychol.* 1980;39(6):1161-1178. doi: https://doi.org/10.1037/h0077714
[26] Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep.* 2024;14(1):6086. doi: https://doi.org/10.1038/s41598-024-06086-9
[27] Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. *Tidymodels.* 2022.
[28] Kuhn M. Building predictive models in R using the caret package. *J Stat Softw.* 2008;28(5):1-26. doi: https://doi.org/10.18637/jss.v028.i05
[29] De Diego IM, Redondo AR, Fernández RR, Navarro J, Moguerza JM. General performance score for classification problems. *Appl Intell.* 2022;52(10):12049-12063. doi: https://doi.org/10.1007/s10489-022-03311-9
[30] Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* 2010;5(9):1315-1316. doi: https://doi.org/10.1097/JTO.0b013e3181ec173d
[31] Badour CL, Resnick HS, Kilpatrick DG. Associations between specific negative emotions and DSM-5 PTSD. *J Interpers Violence.* 2017;32(11):1620-1641. doi: https://doi.org/10.1177/0886260515589930