Decoding the Narrative of Falls: Using Textual Descriptions to Predict Freezing of Gait in Parkinson's Patients

Alireza Rafiei
Department of Computer Science and Informatics, Emory University, Atlanta, GA, USA
(alireza.rafiei@emory.edu)

Abstract

Parkinson's Disease (PD) presents a range of motor challenges for patients, including Freezing of Gait (FOG), a significant contributing factor to falls and subsequent injury. Understanding the early indicators of FOG is crucial for timely intervention. This study aims to leverage Natural Language Processing (NLP) techniques to predict experiencing FOG based on third-person textual narratives describing patient falls. A one-year prospective observational study was conducted with adults previously diagnosed with PD. The dataset consists of various variables, including demographic information, clinical scales, and a narrative of each fall event. We first extracted nine distinct feature sets from the fall narratives. Subsequently, we trained and evaluated seven fine-tuned machine learning classifiers, including a Naive Bayes baseline, on an independent test set. The top-performing model attained an accuracy and micro-averaged F1 score of 0.725, as well as a macro-averaged F1 score of 0.724. An ablation study was also conducted to assess the importance of each feature set and evaluate the optimal training set size for maintaining performance. Our findings indicate that textual descriptions of falls can serve as a predictive marker for FOG in PD patients. This study paves the way for using NLP as a tool in clinical settings for early FOG risk assessment, potentially leading to better patient outcomes.

1 Introduction

Parkinson's Disease (PD) is a progressive neurological disorder that manifests through a range of symptoms, including tremors, rigidity, and bradykinesia (Váradi, 2020). One of the most debilitating motor challenges associated with PD is Freezing of Gait (FOG), characterized by the temporary inability to initiate or continue forward stepping (Snijders et al., 2016). FOG is not only a significant risk factor for falls but also a contributor to the overall decline in the quality of life for PD patients. Consequently, there is an unmet need for reliable tools to predict and assess the risk of FOG, facilitating timely intervention and better patient management.

Traditional assessment methods involve a range of clinical scales, such as the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Hoehn and Yahr Stage. While these scales are informative, they may not capture the full complexity of an individual's lived experience, particularly as it pertains to falls and FOG. Moreover, these assessments often require in-person evaluations, which can be logistically challenging and potentially expose patients to other health risks (van Nimwegen et al., 2013).

In recent years, Natural Language Processing (NLP) techniques have shown promise in healthcare applications, from diagnostic procedures to treatment personalization (Roy et al., 2021). The current study explores the application of NLP in predicting the experience of FOG in PD patients based on textual narratives of their fall incidents. This approach aims to provide a more nuanced understanding of fall events and act as an early warning system for FOG. Of note, we investigate the predictive power of NLP techniques to extract meaningful features from text input data. Specifically, we employ a range of machine learning classifiers to determine their efficacy in predicting FOG, optimize their performance through cross-validation, evaluate their performance on an unseen dataset, and identify key text features indicative of FOG. The overarching goal is to evaluate the viability of employing NLP for early FOG risk assessment in clinical practice.

2 Materials and Methods

2.1 Dataset

The data for this study were collected from a one-year prospective observational study involving 299 adults previously diagnosed with PD by a movement disorders specialist. All participants were recruited and provided written informed consent in accordance with the procedures approved by the Institutional Review Board of Emory University. The participants underwent a single-visit assessment that included a comprehensive cognitive and motor battery aimed at identifying indicators of potential fall risk. Post-enrollment, the participants were prospectively tracked for incident falls over a nominal period of one year. Monthly follow-ups were conducted through mail, email, phone, or text, depending on the participant's preference, to collect information on any falls experienced. Approximately one-third of the enrolled participants reported falls during this period. The dataset encompasses a wide range of variables that fall into three primary categories: demographic information, clinical metrics, and text data. Demographic variables include age, gender, and race, whereas clinical scales include the MDS-UPDRS Part III and the Hoehn and Yahr Stage. Additionally, a third-person textual narrative describes the circumstances and context surrounding each fall event experienced by the participant. The goal is to predict a binary variable indicating whether the participant experiences FOG at baseline. A value of 1 signifies the presence of FOG, and 0 indicates its absence.

2.2 Preprocessing

In preparation for analysis, the textual data underwent a series of preprocessing steps to ensure data quality and consistency. Initially, all text was converted to lowercase to eliminate case sensitivity, thus harmonizing the input. Subsequently, any punctuation marks were stripped from the text, reducing complexity and potential noise. The cleaned text was then tokenized, and common stop words such as 'and,' 'the,' and 'of' were removed from these tokens to focus on the words that carry meaningful information. Lastly, stemming was applied to the tokens to reduce words to their base or root form, and the tokens were put again together to create the textual data for the model development. The word cloud for the preprocessed texts is shown in Figure 1.

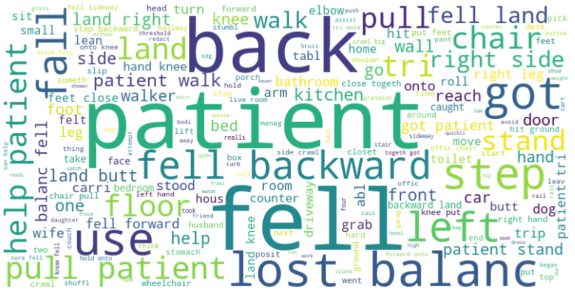

Figure 1: Word cloud of the preprocessed fall description reports.

2.3 Feature Engineering

In the feature engineering stage, nine techniques were employed to extract a rich set of features from the preprocessed text narratives. Initially, we utilized the Count Vectorizer to generate unigram, bigram, and trigram (N-grams of 1 to 3) representations of the text, thus capturing the frequency of specific word combinations in the dataset. Simultaneously, the Term Frequency-Inverse Document Frequency (TF-IDF) method was applied for N-grams ranging from 1 to 3, allowing us to weigh the importance of different terms in the text relative to their occurrence across narratives. These approaches aim to capture the lexical structure and significance of the words in describing fall incidents.

Complementing these traditional NLP methods, we incorporated more advanced techniques to provide a multidimensional view of the text. Word clustering was performed to group similar terms, offering a higher-level abstraction of the textual content. Word2Vec was used to create word embeddings, which capture the semantic context and relationships between words. Sentiment scores were also calculated to gauge the emotional tone conveyed in the fall descriptions. A topic modeling technique, Latent Dirichlet Allocation (LDA), was applied to extract meaningful topics from the text and add the distribution of topics to the feature sets. Also, we developed a bidirectional Long Short-Term Memory (BiLSTM) model for the extraction of pertinent features from the text data. In this regard, the model encompassed a two-layer BiLSTM architecture, each layer containing 64 neurons. This is followed by a dense layer of 16 neurons. The model was trained to distill 16 features from the given textual input. Finally, we employed BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) to extract more complex, contextually rich features from the text. These diverse feature sets were then inputted into multiple machine learning classifiers to assess their predictive power in identifying FOG incidents.

Model	Hyperparameter	Values
LGBM	Learning Rate	{0.01, 0.1}
	Number of estimators	{50, 100, 200, 500}
	Max Depth	{3, 5, 10, 12}
CatBoost	Learning Rate	{0.01, 0.1}
	Iterations	{50, 100, 200, 500}
	Depth	{3, 5, 10, 12}
XGB	Learning Rate	{0.01, 0.1}
	Number of estimators	{50, 100, 200, 500}
	Max Depth	{3, 5, 10, 12}
SVM	C	{0.01, 0.1, 1, 5, 10, 100}
	Kernel	{linear, rbf, poly, sigmoid}
LR	C	{0.01, 0.1, 1, 5, 10, 100}
	Penalty	{none, l2}
	Solver	{liblinear, lbfgs}
RF	Number of estimators	{50, 100, 200, 500}
	Max Depth	{None, 5, 10, 12, 20}
	Min Samples Split	{2, 5}
	Min Samples Leaf	{1, 2, 4}

Table 1: Hyperparameter search space for each classifier.

2.4 Machine Learning Models

Seven machine learning models, along with a Naive Bayes classifier as the baseline, were developed to predict FOG based on the feature-engineered text narratives. Specifically, we utilized Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), and CatBoost as the classifiers. Additionally, we developed a stacking ensemble model that used SVM, LR, and RF as the base models and major voting classifier as the meta-model. Hyperparameter tuning was performed through a grid search approach. This involved exploring a broad pre-defined parameter space for each algorithm (Table 1) and employing 5-fold cross-validation on the training set. The objective was to identify the best-performing model and hyperparameter configuration, as assessed by the overall micro-averaged F1 score. The scripts in this study are available in the corresponding GitHub repository (Assignment 2 repo).

Model	Accuracy	Micro-averaged F1	Macro-averaged F1
NB	0.696	0.696	0.693
LGBM	0.638	0.638	0.638
CatBoost	0.652	0.652	0.639
XGB	0.696	0.696	0.696
SVM	0.710	0.710	0.710
LR	0.725	0.725	0.723
Ensemble	**0.725**	**0.725**	**0.724**
RF	**0.725**	**0.725**	**0.724**

Table 2: Performance metrics of the developed machine learning models on the test set.

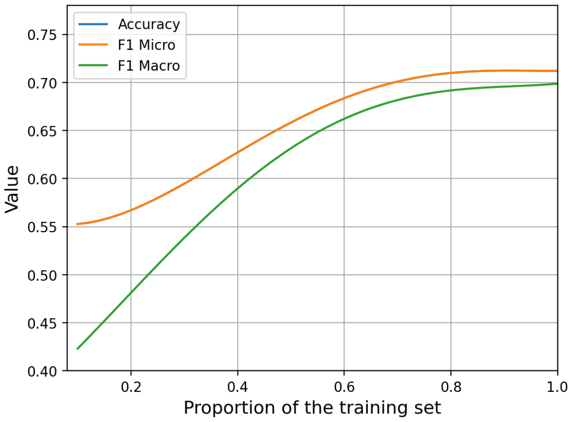

Figure 2: Training set size versus performance graph of the developed RF model.

3 Results

The optimized machine learning models were evaluated using an unseen test dataset containing 71 participant's data. Table 2 summarizes accuracy, micro-averaged, and macro-averaged f1 scores of different machine learning models. As clearly can be seen, the optimized RF model outperformed the other models across all the evaluated performance metrics, achieving an accuracy of 0.725, a micro-averaged F1 score of 0.725, and a macro-averaged F1 score of 0.724. While the baseline classifier trained with all the extracted features had an accuracy of 0.696, two of the developed models, LGBM and CatBoost, lagged behind with accuracies of 0.638 and 0.652, respectively. Figure 2 illustrates how the different evaluated performance metrics changed when varying the training dataset size for the RF model. The trends of interpolation of accuracy and micro-averaged F1 score were approximately similar and marginally superior to the macro-averaged F1 score. The graph reveals that using 80% of the available training data (corresponding to about 239 participants) yields results that are nearly as good as using the entire dataset. This observation suggests that the model's performance improves with an increase in the amount of annotated data, but only up to a saturation point, which in this case is 80% of the data size. Beyond this point, adding more data is unlikely to yield significant improvements in the performance metrics.

3.1 Ablation Study

We perform an ablation study by re-running experiments with removing feature sets. Table 3 provides the optimized RF model trained with different feature sets. For feature set 1, we removed the features extracted from the pre-trained deep learning

Feature set	Accuracy	Micro-averaged F1	Macro-averaged F1
Feature set 1	0.739	0.739	0.730
Feature set 2	**0.754**	**0.754**	**0.743**
Feature set 3	0.667	0.667	0.664
Feature set 4	0.696	0.696	0.689
Feature set 5	0.652	0.652	0.643
Feature set 6	0.681	0.681	0.676
Feature set 7	0.696	0.696	0.689

Table 3: Ablation study evaluating the performance of the top-performing model trained with various feature sets.

models (BERT and BiLSTM) from all the features. Next, the LDA feature set was removed from the feature set 1. Interestingly, by removing those three features, we achieved the highest performance metrics for feature set 2 with an accuracy and micro-averaged F1 score of 0.754 and macro-averaged F1 score of 0.743. Afterward, Word2Vec, sentiment score, text length, and word cluster features were removed to create the feature sets of 3 to 6. Finally, the TF IDF features were dropped out to evaluate the model just with the n-gram features.

We conducted an ablation study in which we iteratively removed various feature sets and re-ran experiments to assess their impact. Table 3 details the performance of the optimized RF model when trained with these modified feature sets. For Feature Set 1, we eliminated the features derived from pre-trained deep learning models, BERT and BiLSTM. Subsequently, we removed the LDA features from Feature Set 1 to create Feature Set 2. Interestingly, this led to the highest performance metrics: an accuracy and micro-averaged F1 score of 0.754 and a macro-averaged F1 score of 0.743. Following this, we successively removed features related to Word2Vec, sentiment scores, text length, and word clusters to generate Feature Sets 3 through 6. Lastly, we excluded the TF-IDF features to evaluate the model's performance solely based on n-gram features.

4 Discussion

The incorporation of NLP techniques for predicting FOG from fall narratives in PD patients was analyzed in this study. The findings indicate that machine learning models can leverage various lexical, semantic, and high-dimensional contextual feature representations extracted from free text descriptions to predict FOG events. While the performance metrics did not reach exceptional levels (for instance, an accuracy rate exceeding 0.95), the analytical pipeline, with achieving balanced and acceptable performance metrics, demonstrated potential for further refinement and development. Additionally, the insights gained from text analysis can be added to other demographics and clinical measures to decipher complex patterns and provide more predictive capabilities.

An ablation analysis revealed that simple n-gram features capture predictive signals on their own, but the integration of additional features derived from more advanced NLP techniques further boosts performance. Having said that, extracting more features, even high-dimensional features derived from deep learning models, does not guarantee to help increase the models' performance. Moreover, the saturation point analysis also provided insights into the minimal dataset size required to attain robust model performance, guiding requirements for future data collection efforts. This analysis also suggested that accumulating more data for training machine learning models in this concept may not directly enhance the performance of the developed models.

Some limitations should be highlighted when interpreting these results. The reliance on self-reported, third-party-described fall incidents may introduce reporting biases. Despite this, the results of the current study represent an important early step toward developing an NLP-powered clinical decision aid for FOG prediction in PD patients. Follow-up studies can focus on extracting more advanced novel features, integrating multimodal data sources, and developing deep learning models. With further refinement, NLP models are expected to show immense promise to enable early preventative interventions for PD patients.

5 Conclusion

This study substantiates the utility of NLP techniques in predicting FOG among PD patients primarily based on textual narratives of fall incidents. Our analysis demonstrates that machine learning classifiers can utilize these textual descriptions as predictive markers for FOG, offering a valuable complement to traditional clinical assessments. While the study does have limitations, such as reliance on self-reported incidents, the findings open avenues for future research to integrate additional features, input data types, and machine learning techniques. Overall, this research can be a promising foundation for employing advanced NLP in early FOG risk assessment, offering a potential pathway toward better patient management and improved healthcare outcomes in PD.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Khushi Roy, Subhra Debdas, Sayantan Kundu, Shalini Chouhan, Shivangi Mohanty, and Biswarup Biswas. 2021. Application of natural language processing in healthcare. *Computational Intelligence and Healthcare Informatics*, pages 393–407.

Anke H Snijders, Kaoru Takakusaki, Bettina Debu, Andres M Lozano, Vibhor Krishna, Alfonso Fasano, Tipu Z Aziz, Stella M Papa, Stewart A Factor, and Mark Hallett. 2016. Physiology of freezing of gait. *Annals of neurology*, 80(5):644–659.

Marlies van Nimwegen, Arlène D Speelman, Sebastiaan Overeem, Bart P van de Warrenburg, Katrijn Smulders, Manon L Dontje, George F Borm, Frank JG Backx, Bastiaan R Bloem, and Marten Munneke. 2013. Promotion of physical activity and fitness in sedentary patients with parkinson's disease: randomised controlled trial. *Bmj*, 346.

Csaba Váradi. 2020. Clinical features of parkinson's disease: The evolution of critical symptoms. *Biology*, 9(5):103.