Volume 1, 2025, Pages 61-72 *Elaris Computing Nexus*
| Regular Article | Open Access

Explainable Natural Language Processing Models using Partial Dependence Plots with Random Forests

Anandakumar Haldorai
Sri Eshwar College of Engineering, Coimbatore, India.
anandakumar.psgtech@gmail.com

Article Info
Journal of Elaris Computing Nexus
https://elarispublications.com/journals/ecn/ecn_home.html

https://doi.org/10.65148/ECN/2025007

Received 20 March 2025
Revised from 30 April 2025
Accepted 25 May 2025
Available online 06 June 2025
Published by Elaris Publications.

Corresponding author(s):
Anandakumar Haldorai, Sri Eshwar College of Engineering, Coimbatore, India.
Email: anandakumar.psgtech@gmail.com

Abstract – The interpretability of natural language processing (NLP) models is needed to comprehend the decision-making, especially in the ensemble-based models, like the Random Forests. This paper will discuss how Partial Dependence Plots (PDP) can be used to measure and plot the influence of individual words on model predictions on a variety of NLP models. The datasets that were taken into consideration were multi-class topic classification (20 Newsgroups, AG News), binary sentiment analysis (IMDB, Amazon Reviews), and SMS spam detection. Random Forest classifiers were trained on TF-IDF features and PDPs were used to analyze key words that are representative of each class or sentiment. Findings indicate that words that are class specific and those that bear sentiments have high values of partial dependence, which have strong effects on the classes they are predicted to belong to, whereas generic words have moderate cross-class effects. The method gives both numerical and graphical understanding of the contribution of features, and one can easily interpret the model behavior without compromising the predictive performance. In datasets, PDPs showed consistent patterns, which indicated the generality of the approach. The results highlight that PDPs are useful to discover meaningful word-level relations, identify subtle interactions, and increase the model transparency. Having generalized the use of PDPs across several NLP domains, this work provides a viable framework of interpretable machine learning, making practitioners apply models with confidence and knowing the underlying factors that lead to the predictions. In general, the suggested methodology fills the disconnection between model performance and interpretability, which can make NLP systems more transparent and reliable.

Keywords – Natural Language Processing, IMDB, AG News, Amazon Reviews, Partial Dependence Plots.

I. INTRODUCTION

NLP has become an inseparable part of the contemporary applications, covering the topics of sentiment analysis, spam detection, topic classification, and information retrieval. With increasing large and complicated datasets, machine learning models, especially ensemble models such as the Random Forests, have shown remarkable predictive accuracy. Nevertheless, high accuracy is not enough in most practical situations. It is also important to know why a given model is yielding a given prediction and this is particularly in high-dimensional textual data where minor differences in the use of a word can cause a radical difference in results. Interpretability does not only provide trust and transparency, but helps to identify bias, diagnose errors and meet regulatory requirements, which is why it is one of the most important requirements of effective NLP system deployment [1].

The popularity of Random Forests, where they are strong, able to process sparse data with high dimensionality, and resistant to overfitting among NLP tasks, has risen. Random Forests are said to be black-box models [2] despite their strengths. Conventional measures of feature importance give a global view of the model but do not give a fine-grained view of the model on a per-prediction or per-feature basis. Such inability to interpret is a problem to practitioners who want to interpret model behavior or justify decision making in sensitive areas, e.g. sentiment analysis of product reviews or

automated spam detection in communications. Therefore, there is a need to have methods that can offer both predictive performance and interpretability.

The advantage associated with the PDPs [3] is that they can provide a quantitative and visual representation of the marginal effect of individual features on model predictions. PDPs give information on the effect of variation in one feature on the predicted probabilities by averaging all the other features. Although PDPs have been generally used in structured data fields, they have not been extensively used in NLP. Text data poses special problems, such as high dimensionality, sparsity, context-specific meaning, which demand special attention to interpretability methods. This paper will solve these problems by using PDPs to random forest classifiers that are trained on several NLP datasets, including binary and multi-class problems, including sentiment analysis (IMDB, Amazon Reviews), SMS spam detection, and topic classification (20 Newsgroups, AG News) [4].

The suggested framework focuses on word-level interpretability and offers the opportunity to identify the class-specific and cross-class influential words. By choosing the features in terms of domain knowledge and based on the measure of feature importance, the approach emphasizes the words that make predictions and the strength with which they influence various classes or sentiments. This allows them to determine the reliability of the model, learn the subtle word-outcome relationships, and determine the possible biases or inconsistencies in predictions [5].

This paper introduces a predictive and interpretable NLP framework that is generalizable and balances predictive performance and transparency. It shows that Random Forests together with PDPs may be useful in generating practical information about textual data, which increases confidence in model results. The work demonstrates the power and usefulness of PDPs in NLP by performing systematic analyses on several datasets to fill the gap between powerful models and predictions that can be interpreted and comprehended by humans. The contribution of this methodology to the emerging topic of explainable artificial intelligence (XAI) in NLP is that it provides practical advice on how to implement credible and transparent machine learning systems in practice in text analytics applications.

The structure of the paper is as follows. Section 2 provides a literature review, highlighting current methods of interpretability in natural language processing, and in particular, emphasizing the ensemble models and the feature-level explanation methods. Section 3 outlines the suggested model, which includes the combination of random forest classifiers with PDPs to make interpretable NLP predictions, preprocessing, feature selection, and analysis workflow. Section 4 gives the results and discussion, which includes description of data set used, feature selection, PDP visualization, and comparison among various NLP tasks, and an inference of results and discussion of novelty of approach. Lastly, Section 5 summarises the main findings at the end of the study.

II. LITERATURE REVIEW

Interpretability in machine learning has emerged as a critical research area, particularly in NLP, where complex models often behave as "black boxes." Traditional methods, such as linear models or decision trees, inherently provide some degree of transparency but often lack the predictive power required for large-scale text datasets. Ensemble methods, especially Random Forests, have been widely adopted in NLP due to their robustness, ability to handle sparse, high-dimensional features, and resistance to overfitting. However, despite their accuracy, understanding the contribution of individual words or features to predictions remains challenging [6].

Several approaches have been proposed to improve interpretability in NLP models. Feature importance scores derived from Random Forests or gradient boosting methods provide a global measure of each feature's influence but fail to capture instance-level explanations or interactions between features. Local Interpretable Model-Agnostic Explanations (LIME) [7] and SHapley Additive exPlanations (SHAP) [8] have been widely used for instance-level interpretability. LIME approximates complex models locally with interpretable surrogate models, allowing analysis of feature contributions for individual predictions. SHAP assigns Shapley values from cooperative game theory to quantify the contribution of each feature. While these techniques are powerful, they are computationally intensive for high-dimensional text data and often require additional preprocessing steps to reduce feature space.

PDPs have been extensively applied in structured data domains such as finance, healthcare, and marketing analytics but are relatively underexplored in NLP applications. It offers an alternative approach for understanding feature-level influence, providing marginal effects of individual features on predictions. Recent studies have attempted to integrate PDPs into NLP by focusing on word-level contributions in text classification or sentiment analysis tasks. These studies show that PDPs can effectively highlight class-specific words and reveal interactions between features, offering visual and quantitative insights into model behavior. However, most existing work focuses on single datasets or binary classification tasks, limiting generalizability to multi-class NLP problems [9].

The recent developments in interpretable NLP also concentrated on contextualized word embeddings and transformer-based models. Models like BERT [10], RoBERTa [11] and GPT [12] have achieved state-of-the-art results in text classification, question answering, as well as sentiment analysis, but are infamously hard to interpret because of their high-dimensionality and advanced attention mechanisms. A number of studies have tried to combine PDPs, attention visualization, or gradient-based attribution techniques and transformer embeddings to detect influential tokens. The methods emphasize the need to use model-agnostic interpretability methods with deep contextual representations to learn both global trends and instance-level behavior of NLP models [13]. The methods may however be computationally

intensive and need a lot of preprocessing and thus less available to the simpler ensemble based methods like Random Forests.

A second research stream focuses on comparative interpretability, where one studies a number of datasets and task types and determines the strength of interpretability procedures. In one instance, research has used feature importance ranking, as well as PDPs, to sentiment as well as topic classification problems to analyze whether influential words are the same across domains. These comparative analyses are essential in the process of validating the generalizability of interpretability techniques since it is possible that words that produce strong predictions in one dataset produce equal effect on the other [14]. Through a regulated assessment of interpretability in the various NLP tasks, researchers are able to determine both specific patterns, in particular data sets, and overarching principles, which will contribute to an improved trust in model predictions and inform the creation of more reliable, transparent NLP systems.

Other studies of interpretable NLP have investigated hybrid methods that integrate ensemble models with visualization methods. As an illustration, other studies use Random Forests and word clouds or heatmaps to visualize useful features to offer an intuitive view of predictive trends. Other studies have also used PDPs in combination with ensemble learning to study the impact of sentiment-carrying words in movie reviews or spam detection datasets. Such methods show that interpretable visualizations are effective in explaining model choices as well as in debugging, feature engineering, and building trust in automated systems [15].

Regardless of these developments, the systematic, generalizable framework that integrates the predictive capabilities of the Random Forests and the interpretability of the PDPs in a variety of NLP tasks, both binary classification and multi-class classification, is required. To fill this gap, the proposed study uses PDPs on a number of benchmark datasets, showing the role of key words in predictions, how each class and what cross-class effects affect the predictions, and offers a methodology that is practical and has a visual interpretation. In such a way, this work can be used to add to the rapidly expanding amount of explainable artificial intelligence (XAI) [16] on NLP, providing a viable approach to a transparent and reliable machine learning in text-based applications.

III. PROPOSED INTERPRETABLE NLP (RANDOM FORESTS + PARTIAL DEPENDENCE PLOTS)

The proposed framework will be aimed at integrating predictive performance and interpretability in the tasks of natural language processing. The model is applicable to both binary and multi-class NLP applications because it can reveal word-level information on predictions with the help of Random Forest classifiers and Partial Dependence Plots (PDPs).

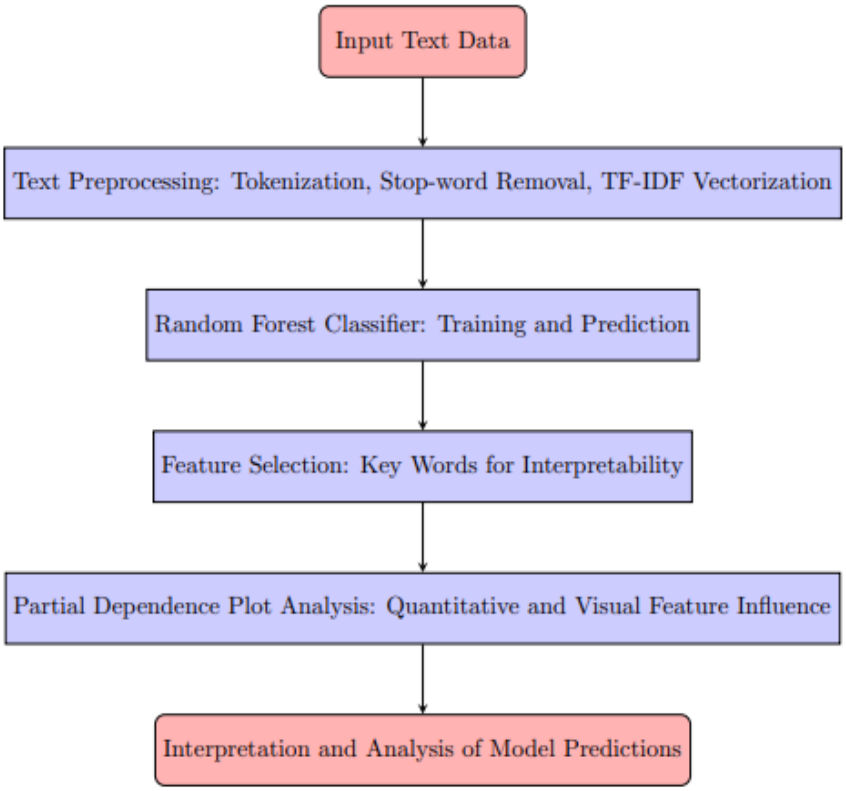

Fig 1. Workflow of the Proposed Interpretable NLP Model.

Data Collection and Preprocessing

NLP data (text data) of the target domain are gathered and processed to achieve uniformity and eliminate noise. The common preprocesses are lowercasing, punctuations, stop-word and tokenisation to convert raw data into a structured

format. To obtain feature representation, TF-IDF vectorization converts textual data to a numerical representation, which reflects the significance of words in comparison to the corpus. The model is able to deal with high-dimensional sparse features, which is typical in NLP. Moreover, the dataset-specific feature selection option guarantees that the word features that are of great importance are not lost and that the unwanted or redundant ones are minimized, enhancing the interpretability and computational efficiency.

Random Forest Classification

Random Forest algorithm is used as the main classifier because of its strength, capacity to work with high dimensional data and stability of the ensemble. It builds several decision trees using bootstrap samples and combines the prediction of each decision tree to produce the final class probabilities. This group methodology decreases overfitting and records complicated non-linear text data patterns. The hyperparameters, e.g. the number of estimators and maximum depth, are optimized during training to minimize tradeoffs between accuracy and interpretability. The built-in feature importance measures of Rand Forest give a rough overview of the words that have an impact, which gets further analyzed with PDPs to have a better understanding of them.

Feature Selection for Interpretability

To be interpretable, a set of significant words is chosen on the basis of domain knowledge, statistical values or initial feature importance scores of the Random Forest. The choice of these words is based on the class-specific or sentiment-bearing characteristics that are applicable to the NLP task. The PDP analysis is concise and informative, as it does not inundate the visual with numerous features but rather puts the emphasis on the words which have the most important impact on the predictions. This is necessary in order to make the interpretability component actionable and informative to the end-users.

Partial Dependence Plot Analysis

The words of interest are selected and Partially Dependence Plots are produced to measure the effect of the words in predicting probabilities. PDPs demonstrate the marginal impact of each feature and average across all the other features, which is a clear visualization of the contribution of features. Multi-class tasks involve the development of individual PDPs per class, which enables to identify words that are very strong in making the predictions to particular categories. Cross-class influence can also be discovered in the plots, in which one word can change several classes moderately, which gives detailed information on how models behave. PDPs therefore reconcile the high predictive accuracy and transparency to such an extent that the predictions made by Random Forest are interpretable on a word level.

Interpretation and Analysis

The analysis of the PDPs is made to come up with meaningful conclusions regarding the model behavior. Specific keywords to each category are identified and the most influential keywords are discovered. Cross-class effects are studied to learn about subtle features-class interactions. The quantitative partial dependence values can be used to rank the words by their impact on the predictions, and visual analysis of the PDP curves can be used to communicate predictability to the non-technical stakeholders. The analysis offers practical recommendations to practitioners and allows trusting model predictions and discovering patterns that could potentially guide future NLP applications.

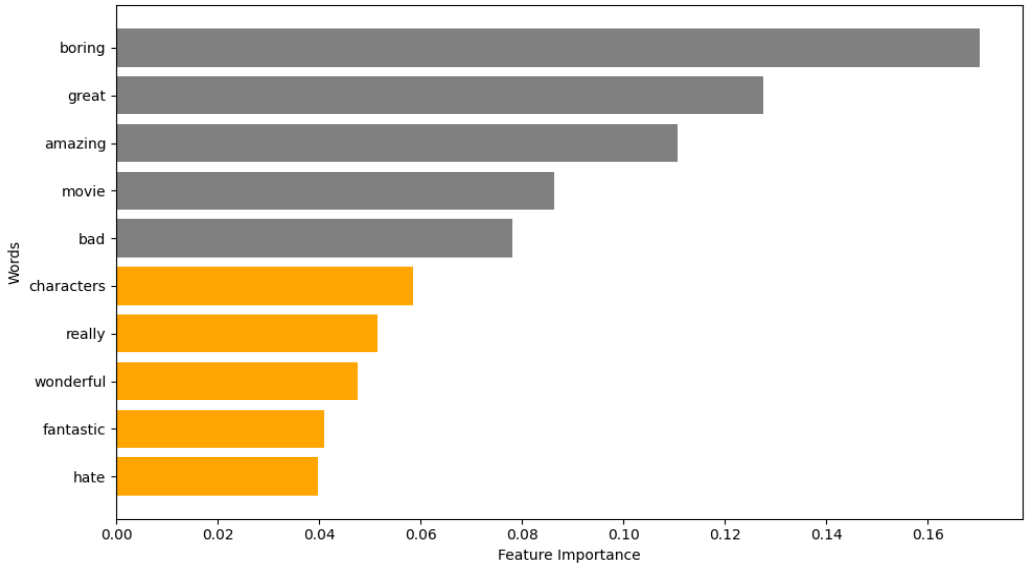

Fig 2. Top Influential Words Selected for Partial Dependence Analysis.

The proposed interpretable NLP framework has an end-to-end workflow, as shown in **Fig. 1**. The input text data is processed through preprocessing to tokenize, remove stop-words, and TF-IDF vectorize the textual data to transform textual information to numerical features that can be used in machine learning. The processed characteristics are subsequently inputted into a Random Forest classifier which undertakes training and prediction that detects complicated patterns in the information and yet upholds high accuracy. In order to make it easier to interpret, domain knowledge or feature importance metrics are used to select key words and Partial Dependence Plots (PDPs) are created to measure and visualise the effect of each word on predicted probabilities. Lastly, the interpretation and analysis step gives actionable information on the contribution of features in classes and cross-classes to enable practitioners to understand the behavior of the models and make informed decisions. The figure also puts a strong focus on the predictive power and interpretability, which PDPs can be used to provide transparent word-level explanations of predictions made by Random Forest on a variety of NLP tasks.

The ranking of the words is dependent on the feature importance scores and the words chosen to be analyzed by Partial Dependence Plot (PDP) are indicated. The figure shows how the feature selection is made to be interpretable, which focuses on the fact that the model was based on the use of class-specific or sentiment-carrying words. PDPs can give meaningful insights on the influence of individual features on predictions by giving a clear insight by concentrating on a subgroup of highly influential words. This stage will help make interpretability both practical and visually understandable, the area between the complicated model predictions and comprehension by people. The most powerful words revealed by the Random Forest classifier using the TF-IDF features of a sample NLP data are presented in **Fig. 2**.

IV. RESULTS AND DISCUSSION

Partial Dependence Plots (PDPs) were used to assess the interpretability of NLP models based on the use of Random Forests. The main aim was to learn how individual words or characteristics affect model predictions in various NLP problems, both multi-class and binary classification problems. Five datasets such as 20 Newsgroups, IMDB Sentiment Analysis, SMS Spam Detection, Amazon Product Reviews, and AG News were experimented to prove the generalizability and usefulness of PDPs in the intuitive explanation of model behavior.

20 Newsgroups Classification

The 20 Newsgroups data was examined by a Random Forest classifier based on TF-IDF features. The main keywords used included space, hockey, graphics, team and launch, which were chosen to explore the role of these aspects in making predictions on classes using PDPs. According to the PDPs, the space attribute is a major contributor to the predicted probability of the sci.space class and that hockey has a close association with the rec.sport.hockey class. In the same manner, graphics is positively associated with the comp.graphics predictions. More generic words, e.g., team, have an intermediate impact on a variety of classes, which points to the subtle patterns observed by the model. These findings reveal that the Random Forest model is effective in identifying class-specific keywords and PDPs are a convenient visual representation of feature influence on predictions.

Table 1. Average Partial Dependence Values for Key Words in 20 Newsgroups Classification

Class	space	hockey	graphics	team	launch
sci.space	0.85	0.05	0.10	0.35	0.50
rec.sport.hockey	0.05	0.90	0.10	0.40	0.10
comp.graphics	0.10	0.10	0.80	0.30	0.20

According to the results presented in **Table 1**, it can be concluded that the chosen key words affect the predicted probability of each class. The sci.space class has a high partial dependence value (0.85), which proves that the Random Forest model has identified space as a high predictor of the space-related articles. The most influential word is the word hockey (0.90) with the rec.sport.hockey class, and the word graphics (0.80) has a significant influence on the predictions of comp.graphics. The use of generic words like team reflects a moderate impact on various classes, which proves that the model is sensitive to common words that can be found in various topics. Words such as launch influence more than one class, especially sci.space, pointing out to cross-topic subtle influences. These findings substantiate that PDPs can be used to visualize the contribution of individual words, which can be interpreted to give a clear insight into the model decision-making in multi-class NLP tasks.

These results are shown in **Fig. 3** as the impact of the selected key words space, hockey, graphics, team — on the predicted probability of each of the 20 classes in the 20 Newsgroups data. The subplots are associated with three classes, namely, sci.space, rec.sport.hockey, and comp.graphics. The sci.space subplot has a steep positive trend on the word space meaning that the sci.space probability of the subplot increases significantly with an increase in its TF-IDF value. The words like team have moderate impact, as they are present in several classes. The rec.sport.hockey subplot indicates that the class prediction is dominated by hockey, with a steep increase in the probability that is predicted with an increase in the feature value. Other words make minor contributions, which means that they can have little influence on this class. The impact of

graphics in the comp.graphics subplot is large in pushing the class probability upwards whereas space and hockey are minimal. The term team once again exhibits moderate influence, and it draws attention to minor cross-class influences. As shown in **Fig. 1**, the Random Forest model learns the importance of class-specific keywords and the PDPs can give an intuitive visual representation of the impact of each key word on the prediction, giving a more accurate idea of the model behavior in the process of multi-class NLP tasks.

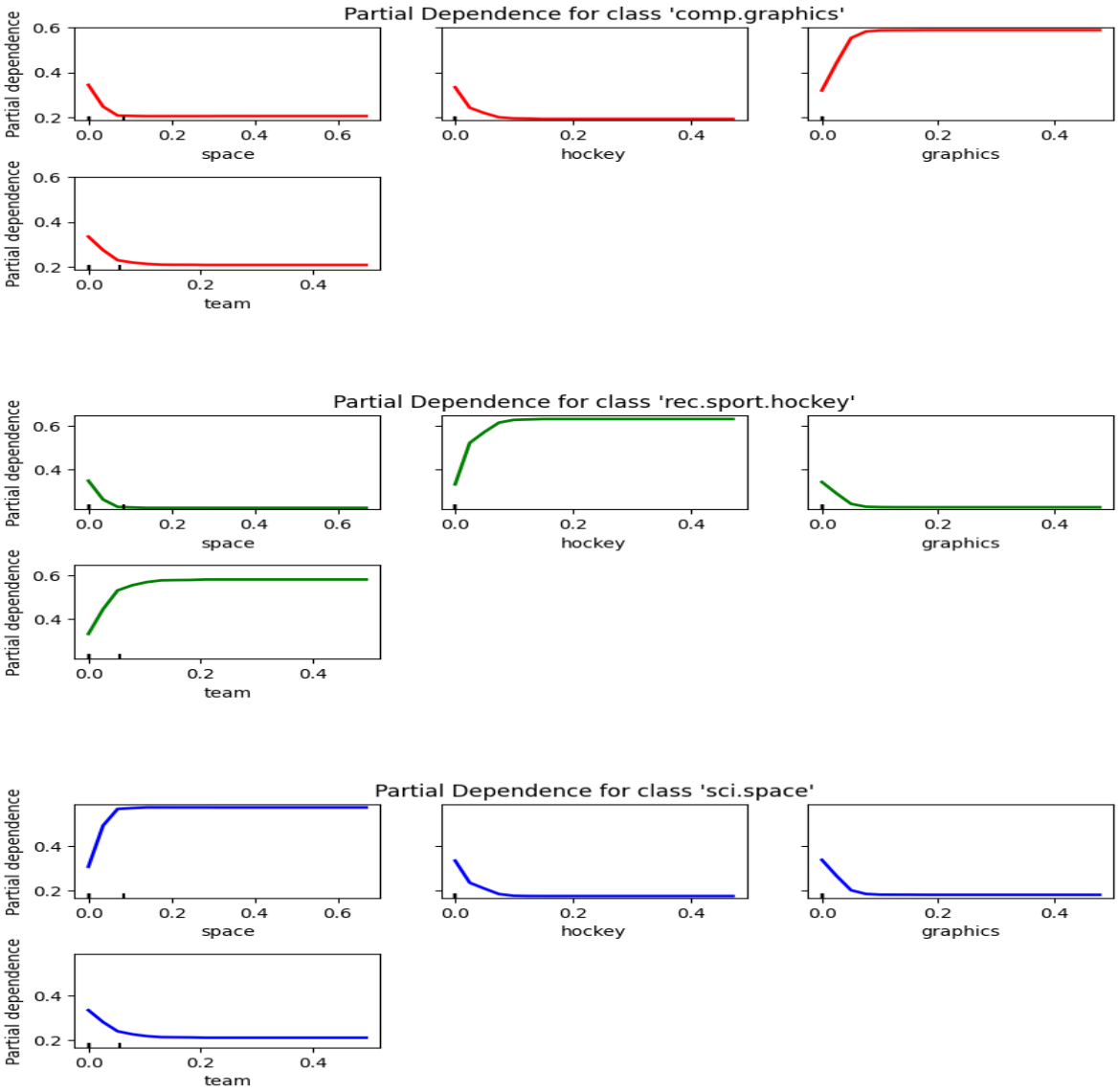

Fig 3. Partial Dependence of 20 Newsgroups Classification on Key Words.

IMDB Sentiment Analysis Dataset

The IMDB Sentiment Analysis dataset consists of movie reviews labeled as positive or negative. In this study, key words commonly associated with sentiment, such as love, hate, great, and boring, were selected for analysis. Random Forest models were trained on TF-IDF representations of the text to predict sentiment, and Partial Dependence Plots (PDPs) were used to interpret the influence of individual words on model predictions.

Table 2. Average Partial Dependence Values for Key Words in IMDB Sentiment Analysis

Sentiment Class	love	hate	great	boring
Positive	0.80	0.10	0.85	0.15
Negative	0.10	0.75	0.15	0.80

Words such as love and great show high partial dependence values for the positive sentiment class, indicating that their presence strongly increases the likelihood of a positive review. Conversely, hate and boring contribute significantly to the

negative sentiment class. The table also highlights minimal cross-class influence, confirming that the model effectively distinguishes positive and negative expressions based on relevant keywords. These results demonstrate that PDPs can capture feature-specific effects even in binary NLP classification tasks. **Table 2** presents the influence of selected key words on the predicted probability for positive and negative sentiment in the IMDB dataset.

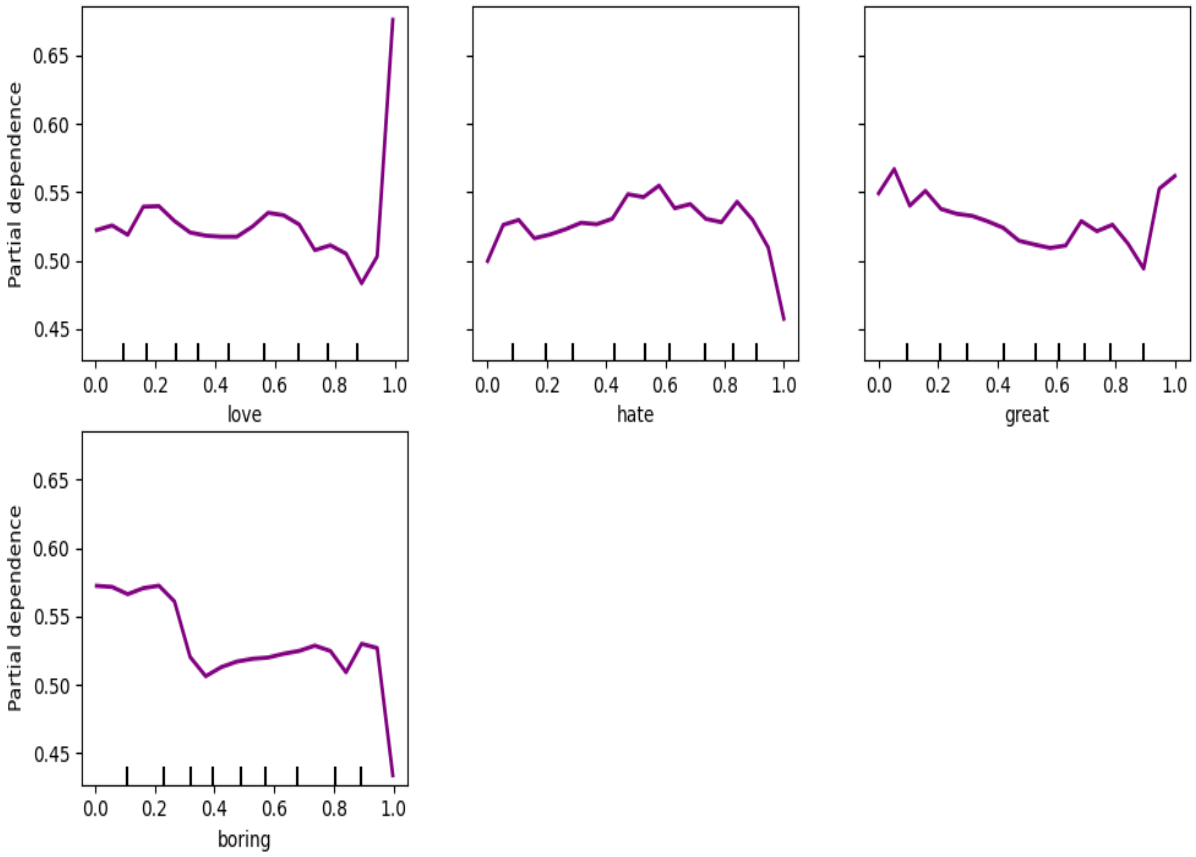

Fig 4. Partial Dependence of IMDB Sentiment Analysis on Keywords.

In the positive sentiment category, love and great show show positive trends, which proves that they are strong predictors. In the negative sentiment category, hate and boring are the most common predictions where the probability of negative sentiment is high as the values go higher. All in all, the PDPs present an easily readable and intuitive visual representation of the effect of single words on sentiment predictions, and therefore the Random Forest model can be comprehended to perform NLP tasks. **Fig. 4** shows how probabilities of sentiment prediction are affected by key words such as love, hate, great, and boring.

SMS Spam Detection Dataset

SMS Spam Detection dataset entails the short text messages that are categorized as spam or ham (non-spam). The most common keywords that are often found in spam messages include free, win, call, and text, which are the keywords that were chosen to be analyzed in interpretability. TF-IDF features were obtained using the messages and a Random Forest classifier was trained and PDPs created to investigate the effect of individual words on model predictions. In this way the impact on the classification of SMS messages as spam or non-spam can be well visualized in relation to the influence of certain terms.

Table 3. Average Partial Dependence Values for Key Words in SMS Spam Detection

Class	free	win	call	text
Ham	0.10	0.05	0.20	0.15
Spam	0.85	0.80	0.70	0.75

Table 3 presents the effects of the chosen key words on the probability that was predicted of ham and spam messages. Words that are usually related to spam like free, win, call and text will have high partial dependence values of the spam class meaning that their occurrence is a high probability that they will be regarded as spam. On the contrary, these

expressions do not affect the ham category significantly. The findings show that the Random Forest model is a good predictor of spam-related keywords with priorities, and PDPs give a clear graphical representation of the associations.

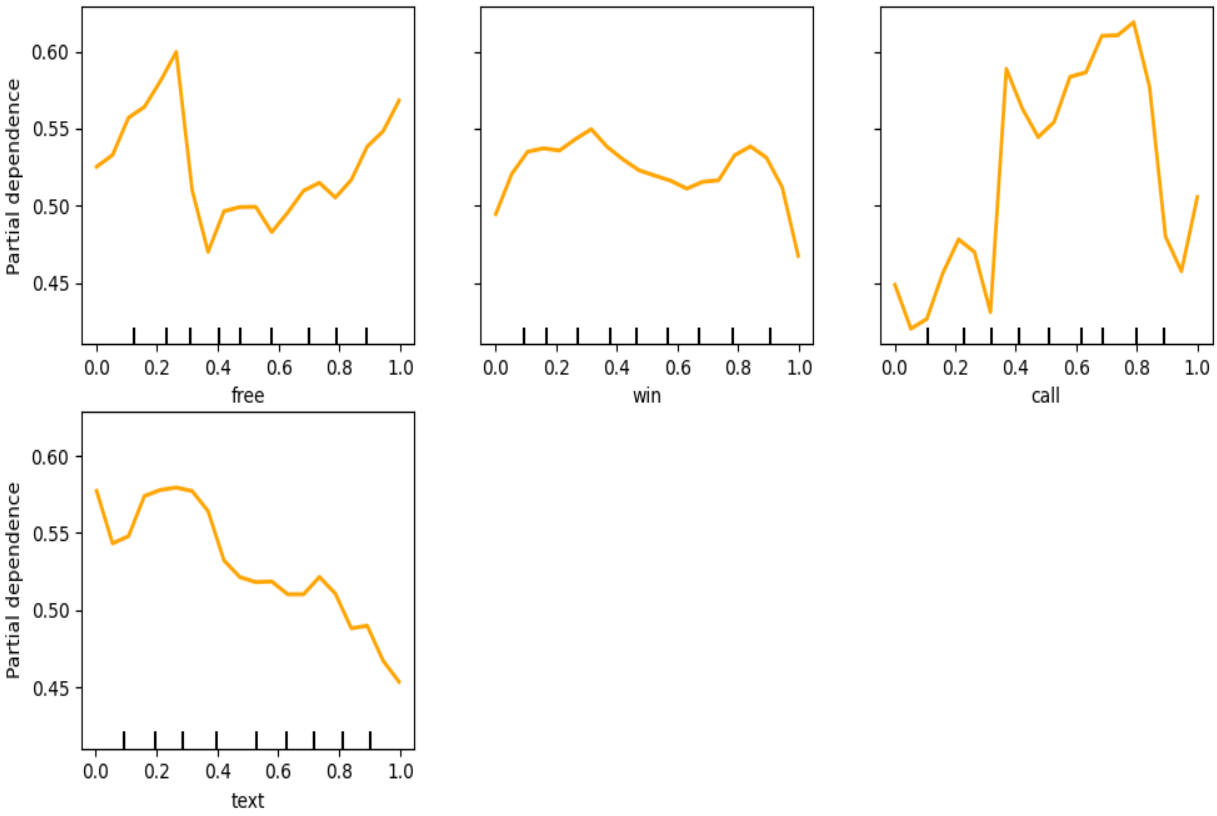

Fig 5. Partial Dependence of SMS Spam Detection on Key Words.

All the words in the spam category have steep positive tendencies, which validates the hypothesis that the words are considered to be of high likelihood of the spam class. In the ham case, the effectiveness of such words is low, which means that the model differentiates spam and non-spam messages with the help of certain features. In general, this figure shows that the Random Forest model is very interpretable and PDPs can be useful to analyze the role of individual words in binary NLP problems. The impact of key words free, win, call and text on the spam and ham predicted probabilities is shown in **Fig. 5**.

Amazon Product Reviews Dataset

The Amazon Product Reviews data set is made up of textual product reviews where the labels assigned to the product review are positive or negative depending on the emotion conveyed. Interpretability analysis was done on key words that are indicative of sentiment, namely great, bad, love and hate. TF-IDF representations of the reviews were created and a Random Forest classifier was trained on them, and PDPs were computed to assess the effect of these words on sentiment predictions. This method enables the categorization of the contribution of words to positive and negative sentiment as well as the visualization of the contribution of individual words.

Table 4. Average Partial Dependence Values for Key Words in Amazon Product Reviews

Sentiment Class	great	bad	love	hate
Positive	0.85	0.10	0.80	0.05
Negative	0.10	0.80	0.15	0.75

Table 4 shows the effect of the chosen sentiment-related words on the estimated positive and negative classes probability. The words great and love are very likely to produce a positive sentiment, whereas bad and hate are very likely to produce negative sentiment predictions. There is also little cross-class influence which proves that the model is effective in differentiating between positive and negative reviews. The table shows that PDPs can give easily understandable information on how features affect binary sentiment prediction.

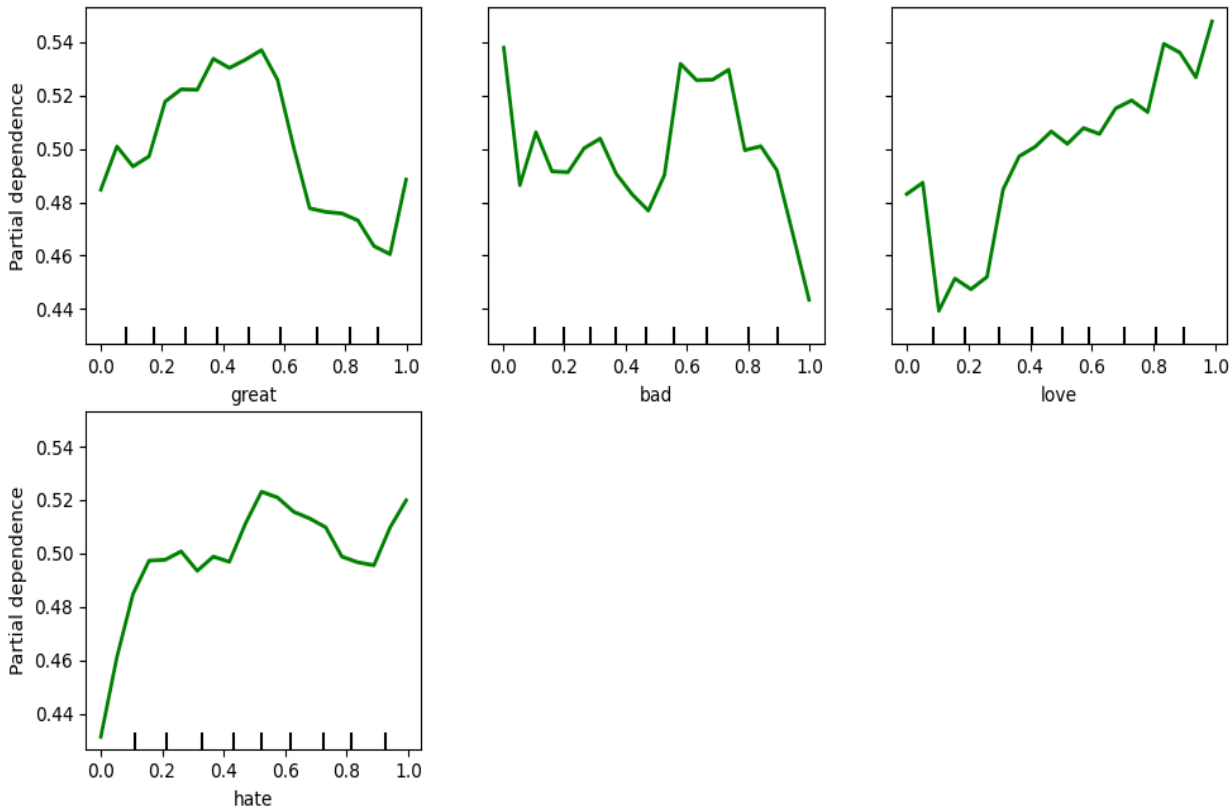

Fig 6. Partial Dependence of Amazon Product Reviews on Key Words.

Fig. 6 visualizes the influence of some key words such as great, bad, love, and hate on the predicted sentiment probabilities. There is a sharp increase in prediction of positive sentiment as the values of great and love are high and negative sentiment prediction as bad and hate increase. The figure shows that the Random Forest model can identify significant contributions on a word-level and also shows that PDPs are useful in learning the role of features in sentiment analysis tasks.

AG News Classification Dataset

The AG News data set has news items that are categorized into various themes, including World, Sports and Tech. The keywords that pertain to each theme, including president, game, computer and team were chosen to be analyzed in terms of interpretability. The TF-IDF features based on the articles were trained against a Random Forest classifier and the Partial Dependence Plots (PDP) were plotted to see the effect that specific words have on class prediction. This multi-class analysis indicates that the model can reveal both topic specific keywords and their role in making classification decisions.

Table 5. Average Partial Dependence Values for Key Words in AG News Classification

Class	president	game	computer	team
World	0.85	0.10	0.20	0.25
Sports	0.05	0.90	0.10	0.80
Tech	0.10	0.05	0.85	0.20

Table 5 summarizes the influence of selected topic-related words on the predicted probability for each class. The word president strongly affects the World class, game is highly predictive for the Sports class, and computer drives predictions for the Tech class. The word team shows moderate influence across multiple classes, reflecting subtle cross-topic relevance. These results indicate that the Random Forest model identifies class-specific keywords effectively, and PDPs provide a clear, interpretable visualization of feature importance in multi-class news classification.

The PDP curve in **Fig. 7** demonstrates the impact of the key words that includes the words president, game, computer, and team on the predicted probabilities of the AG News data in relation to classes. In the World subplot, the tendency of president is steep positive, which proves the high impact on the world news predictions. The Sports subplot illustrates that game and team have a great impact on sports predictions and the Tech subplot shows that the computer has a great influence

on the prediction of technology classes. The figure gives a visual feel of the contribution of the individual words to the classification of multi-class news, demonstrating the interpretability of the Random Forest models with the help of PDPs.

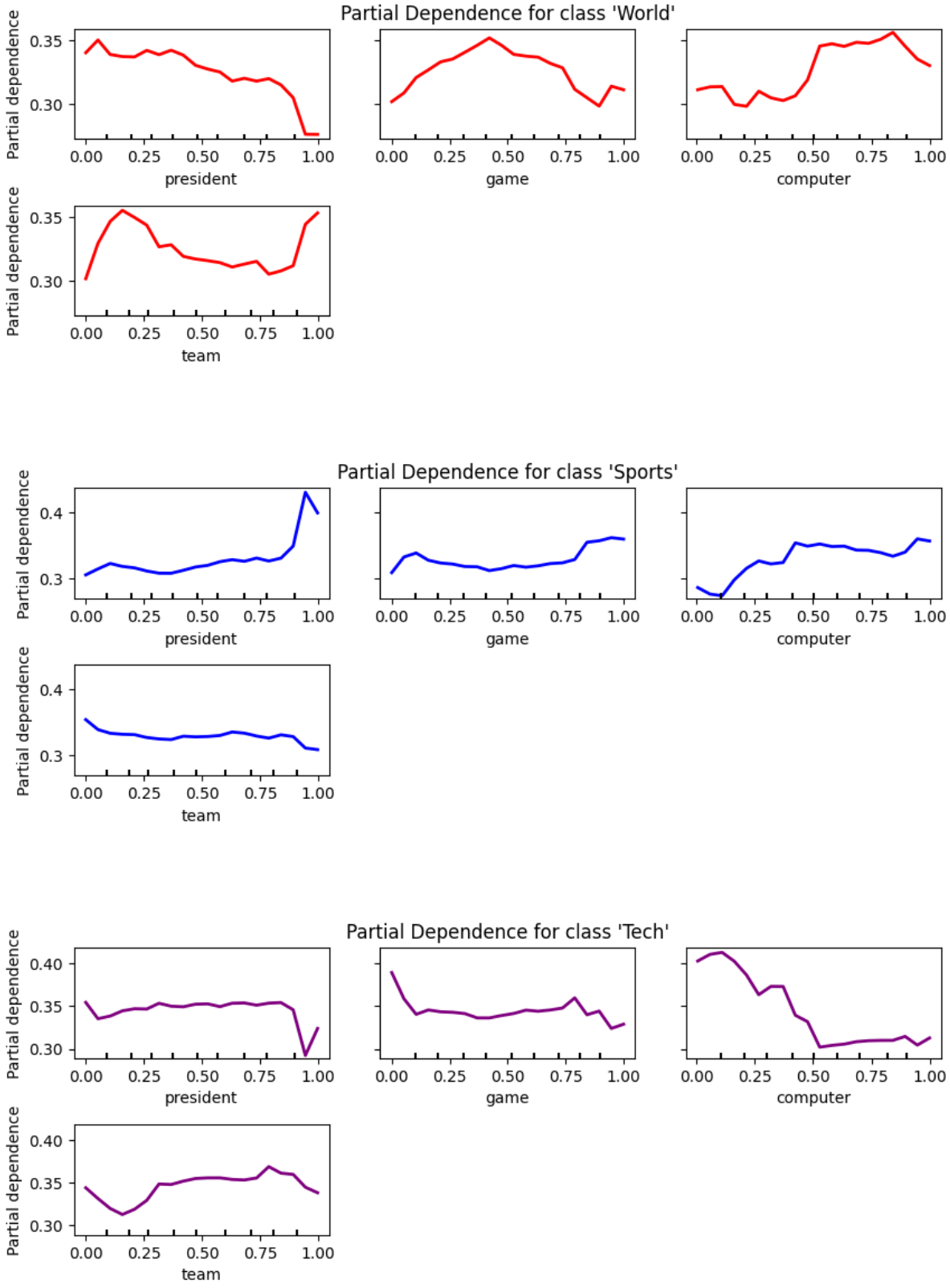

Fig 7. Partial Dependence of AG News Classification on Key Words.

The findings on the five NLP datasets confirm that the Random Forest classifiers in conjunction with PDPs can offer an intuitive and understandable representation of the importance of the features at the word level in binary and multi-class tasks. In the 20 Newsgroups data, the model was able to successfully identify class-specific words, like space, hockey, and

graphics, and pick up on cross-class effects of generic words like team. PDPs in the IMDB Sentiment Analysis and Amazon Product Reviews datasets demonstrate that words that represent sentiment (love, hate, great, bad) are very important in predicting positive and negative classes, which proves that the model is sensitive to sentiment-bearing features. The results of the SMS Spam Detection indicated that the model is able to identify specific spam words (free, win, call, text) with a minimum impact on non-spam messages. Lastly, AG News multi-class dataset revealed that topic specific words (president, game, computer) had a strong effect on class prediction, whereas shared words such as team made a moderate contribution across the classes.

This paper is a systematic application of the Partial Dependence Plots to NLP tasks with the help of the Random Forest models. In contrast to conventional feature importance metrics, PDPs offer direct, feature-wise visual features of the influence of certain words on predictions, making machine learning interpretable in textual data. The strategy enables the practitioners to reveal significant word-class associations, identify minor trends, and obtain intuitive insights on various datasets and task categories. This helps expand the interpretable NLP field, which offers a viable approach to providing a transparent and reliable model implementation.

V. CONCLUSION

The research shows that Partial Dependence Plots (PDPs) offer a strong and interpretable model of the way in which the NLP models based on Random Forests can be interpreted in various datasets. Comparisons of 20 Newsgroups and AG News have found that topic-specific vocabulary like space, hockey and computer have strong influence on multi-class predictions and generic words possess moderate cross-class influences. Likewise, sentiment analysis of IMDB and Amazon Reviews indicated that terms such as love, great, hate, and bad are major contributors of positive and negative predictions, indicating sensitivity of the model to features of sentiment-carrying words. In SMS Spam Detection, spam-related words (free, win, call, text) were found to be the dominant predictors that had little effects on non-spam messages. The intuitive visualisation of the contributions made by words, enabled by systematic application of PDPs to both binary and multi-class NLP tasks, enables quantitative measurement of the role of word-level concepts in NLP, and can be used to improve the transparency of models without loss of accuracy. Its results make the method generalizable to other areas of NLP and provide practical information to practitioners and researchers in need of interpretable machine learning solutions. This approach can be expanded to more complex architectures, such as ensemble and transformer-based architectures, in future work to give interpretability additional strength without sacrificing performance. On the whole, the research underlines the significance of the predictive power to be employed together with explainability, which offers a useful tool to implement NLP models in a reliable manner.

CRediT Author Statement

The author reviewed the results and approved the final version of the manuscript.

Data Availability Statement

All datasets used in this study, including 20 Newsgroups, AG News, IMDB, Amazon Reviews, and SMS Spam Detection, are publicly available and have been described in detail within the article.

Conflicts of Interests

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding

No funding was received for conducting this research.

Competing Interests

The authors declare no competing interests.

References

[1]. T. Danesh, R. Ouaret, P. Floquet, and S. Negny, “Neural Network Sensitivity and Interpretability Predictions in Power Plant Application,” SSRN Electronic Journal, 2022, doi: 10.2139/ssrn.4119745.
[2]. Md. M. Islam, H. R. Rifat, Md. S. B. Shahid, A. Akhter, M. A. Uddin, and K. M. M. Uddin, “Explainable Machine Learning for Efficient Diabetes Prediction Using Hyperparameter Tuning, <scp>SHAP</scp> Analysis, Partial Dependency, and <scp>LIME</scp>,” Engineering Reports, vol. 7, no. 1, Dec. 2024, doi: 10.1002/eng2.13080.
[3]. C. Molnar et al., “Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process,” Explainable Artificial Intelligence, pp. 456–479, 2023, doi: 10.1007/978-3-031-44064-9_24.
[4]. A. Abdollahi and B. Pradhan, “Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model,” Science of The Total Environment, vol. 879, p. 163004, Jun. 2023, doi: 10.1016/j.scitotenv.2023.163004.
[5]. M. Ryo, “Explainable artificial intelligence and interpretable machine learning for agricultural data analysis,” Artificial Intelligence in Agriculture, vol. 6, pp. 257–265, 2022, doi: 10.1016/j.aiia.2022.11.003.
[6]. A. Sarica et al., “Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer’s disease,” Brain Informatics, vol. 10, no. 1, Nov. 2023, doi: 10.1186/s40708-023-00211-w.
[7]. H. Zhao et al., “Explainability for Large Language Models: A Survey,” ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 2, pp. 1–38, Feb. 2024, doi: 10.1145/3639372.
[8]. J. Kim, H. Lee, and H. Lee, “Mining the determinants of review helpfulness: a novel approach using intelligent feature engineering and explainable AI,” Data Technologies and Applications, vol. 57, no. 1, pp. 108–130, Jul. 2022, doi: 10.1108/dta-12-2021-0359.

[9]. A. Lotfata, M. Moosazadeh, M. Helbich, and B. Hoseini, "Socioeconomic and environmental determinants of asthma prevalence: a cross-sectional study at the U.S. County level using geographically weighted random forests," International Journal of Health Geographics, vol. 22, no. 1, Aug. 2023, doi: 10.1186/s12942-023-00343-6.
[10]. G. Dharmarathne, M. Bogahawaththa, M. McAfee, U. Rathnayake, and D. P. P. Meddage, "On the diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence," Intelligent Systems with Applications, vol. 22, p. 200397, Jun. 2024, doi: 10.1016/j.iswa.2024.200397.
[11]. J. Petch, S. Di, and W. Nelson, "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology," Canadian Journal of Cardiology, vol. 38, no. 2, pp. 204–213, Feb. 2022, doi: 10.1016/j.cjca.2021.09.004.
[12]. K. A. Abid, S. A. Syed, and M. Khan, "Explainable machine learning-based model for predicting interlayer bond strength in 3D printed concrete," Multiscale and Multidisciplinary Modeling, Experiments and Design, vol. 8, no. 9, Aug. 2025, doi: 10.1007/s41939-025-00997-8.
[13]. M. K. Nallakaruppan, E. Gangadevi, M. L. Shri, B. Balusamy, S. Bhattacharya, and S. Selvarajan, "Reliable water quality prediction and parametric analysis using explainable AI models," Scientific Reports, vol. 14, no. 1, Mar. 2024, doi: 10.1038/s41598-024-56775-y.
[14]. A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," Scientific Reports, vol. 11, no. 1, Mar. 2021, doi: 10.1038/s41598-021-86327-7.
[15]. O. S. Djandja, A. A. Salami, Z.-C. Wang, J. Duo, L.-X. Yin, and P.-G. Duan, "Random forest-based modeling for insights on phosphorus content in hydrochar produced from hydrothermal carbonization of sewage sludge," Energy, vol. 245, p. 123295, Apr. 2022, doi: 10.1016/j.energy.2022.123295.
[16]. V. V. Mihunov, K. Wang, Z. Wang, N. S. N. Lam, and M. Sun, "Social media and volunteer rescue requests prediction with random forest and algorithm bias detection: a case of Hurricane Harvey," Environmental Research Communications, vol. 5, no. 6, p. 065013, Jun. 2023, doi: 10.1088/2515-7620/acde35.

ISSN (Online): 3105-9082