Literature Review: Twitter Sentiment Classification

Jeffrey Clancy
Esteban Felix Tapia
Stepan Ochodek
Alen Pavlovic

Master of Science in Applied Data Science
University of Chicago

Course: Linear and Nonlinear Models for Business Application
Professor: Dr. Utku Pamuksuz
TA: Irem Pamuksuz

August 3, 2024

1 Literature Review: Sentiment Analysis in Twitter Using Machine Learning Techniques

1.1 Research Paper Overview

The paper "Sentiment Analysis in Twitter Using Machine Learning Techniques" by Neethu M S and Rajasree R explores the challenges and methodologies for analyzing sentiments in tweets about electronic products. It emphasizes the difficulties posed by the informal language and brevity of tweets.

1.2 Problem Covered in the Paper

The main challenge addressed is the difficulty of sentiment analysis on Twitter due to short tweet lengths, slang, misspellings, and emoticons. The goal is to accurately classify the sentiment of tweets as positive or negative despite these issues.

1.3 Proposed Models for Solutions

The paper proposes a solution involving multiple steps:

- **Preprocessing**: This step includes removing URLs, handling slang, and correcting misspellings.
- **Feature Extraction**: Conducted in two phases. Initially, Twitter-specific features like hashtags and emoticons are extracted. These features are then removed from the tweets to perform standard text feature extraction.
- **Classification Models**: Several machine learning classifiers are employed, including:
  - **Naive Bayes (NB)**
  - **Support Vector Machine (SVM)**
  - **Maximum Entropy (ME)**
  - **Ensemble Classifier**: Combines NB, SVM, and ME using a voting mechanism.

1.4 Comparison with Our Project Topic

Our project, "Twitter Sentiment Classification," shares a similar focus on classifying user sentiment in tweets. The model in the paper is also designed to classify sentiment in longer form reviews. Both projects will evaluate models based on precision, recall, F1 score, and accuracy, with these metrics calculated separately for positive and negative sentiments.

1.5 Adaptability of Proposed Model

- **Preprocessing Steps**: Integrating the preprocessing techniques for handling slang and misspellings can improve our data quality.
- **Feature Vector Creation**: Including emoticons, hashtags, and part-of-speech tags in our feature vector can capture syntactic and semantic nuances, enhancing classification performance.

1.6 Conclusion

The methodologies and findings of this paper offer a robust framework that can be adapted to our project. By incorporating their preprocessing techniques, feature extraction methods, and classification models, we can aim for higher accuracy and more reliable sentiment analysis.

1.7 Reference

Neethu M S, Rajasree R, "Sentiment Analysis in Twitter Using Machine Learning Techniques," 4th ICCCNT 2013, IEEE, 2013.

2 Literature Review: ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for Sentiment Analysis

2.1 Research Paper Overview

The paper "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for Sentiment Analysis" by Basiri et al. addresses the limitations of various popular sentiment analysis models by combining the strengths of CNNs and RNNs with attention mechanisms to improve accuracy and comprehensiveness in sentiment analysis.

2.2 Problem Covered in the Paper

The primary challenge addressed in this paper is overcoming the limitations of individual sentiment analysis models like deep neural networks, long short-term memory (LSTM), and gated recurrent units (GRU). While each of these models has its strengths, they also have significant limitations that this paper aims to mitigate.

2.3 Proposed Models for Solutions

The proposed solution uses several advanced techniques:

- **CNNs and RNNs**: Capture important words and phrases and identify their relationships over longer passages.
- **Attention Mechanism**: Understands the context and semantics of the text.
- **Bidirectional Processing**: Provides enhanced context by considering words before and after the target word.
- **Deep Architecture**: Incorporates multiple CNNs and RNNs for greater data processing and comprehension.

These combined techniques result in higher accuracy across various classification performance metrics compared to six competing models.

2.4 Comparison with Our Project Topic

Our project, "Twitter Sentiment Classification," shares a similar focus on classifying user sentiment in tweets. The model in the paper is also designed to classify sentiment in longer form reviews. Both projects will evaluate models based on precision, recall, F1 score, and accuracy, with these metrics calculated separately for positive and negative sentiments.

2.5 Relevance to Our Project

The model created by Basiri et al. highlights valuable techniques that can be explored in our project. It demonstrates the effectiveness of combining multiple modeling techniques to improve performance rather than them interfering with each other.

2.6 Adaptability of Proposed Model

If time permits, our team will consider incorporating some of the techniques highlighted in this paper. It is worth noting that the model was developed to work on both short-form text (tweets) and long-form text (reviews). Some advanced techniques may not significantly impact sentiment classification exclusively for tweets.

2.7 Conclusion

The methodologies and findings of this paper provide a robust framework that can be adapted to enhance our project's accuracy and reliability. By incorporating their advanced modeling techniques, we can potentially achieve better performance in sentiment classification.

2.8 Reference

Basiri, Mohammad Ehsan, et al. "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for Sentiment Analysis." Expert Systems with Applications 149 (2020): 113240.

3 Literature Review: Sentiment Analysis in the Age of Generative AI

3.1 Research Paper Overview

The paper "Sentiment Analysis in the Age of Generative AI" by Jan Ole Krugmann and Jochen Hartmann explores the performance of state-of-the-art large language models (LLMs) such as GPT-3.5, GPT-4, and Llama 2 in zero-shot sentiment classification tasks. It addresses the challenge of leveraging generative AI models for sentiment analysis without the need for extensive task-specific training.

3.2 Problem Covered in the Paper

This paper investigates the performance of LLMs in zero-shot sentiment classification, highlighting the difficulties of using generative AI models for sentiment analysis tasks without extensive task-specific training.

3.3 Proposed Models for Solutions

The study benchmarks LLMs against traditional transfer learning models like SiEBERT and fine-tuned RoBERTa. Despite their zero-shot nature, LLMs can compete with and sometimes surpass traditional models in terms of sentiment classification accuracy. The analysis also examines the influence of textual data characteristics, such as text complexity and the presence of lengthy content-laden words, on classification accuracy.

3.4 Comparison with Our Project Topic

Our project on Twitter sentiment analysis using the Sentiment140 dataset from Kaggle shares the goal of classifying sentiments in social media data. Both projects seek to improve sentiment classification accuracy and handle the intricacies of social media text.

3.5 Relevance to Our Project

The insights from this paper are highly relevant as they demonstrate the capabilities of LLMs in sentiment analysis. The use of LLMs can significantly enhance the robustness and accuracy of sentiment classification in our project. Additionally, the zero-shot learning capability of LLMs can be particularly beneficial for our project, allowing us to handle diverse and evolving social media data without the need for extensive retraining.

3.6 Adaptability of Proposed Model to Our Project

Integrating LLMs like GPT-4 and Llama 2 into our Twitter sentiment analysis project could simplify the process of adapting the model to new types of text data by leveraging the pre-trained capabilities of these models. The ability to perform sentiment classification without extensive task-specific training aligns with our need to efficiently manage a wide range of text data from social media platforms like Twitter. Moreover, the paper's focus on data characteristics such as text complexity and structured content suggests that we should consider enhancing our preprocessing steps to improve model performance.

3.7 Conclusion

The methodologies and findings of this paper provide valuable insights into leveraging advanced LLMs for sentiment analysis. By incorporating LLMs like GPT-4 and Llama 2, our project can potentially achieve higher accuracy and robustness in sentiment classification. The zero-shot learning capabilities of these models can also streamline the process of handling diverse and evolving text data from social media.

3.8 Reference

Krugmann, Jan Ole, and Jochen Hartmann. "Sentiment Analysis in the Age of Generative AI." Customer Needs and Solutions (2024): 1-19. doi:10.1007/s40547-024-00143-4.

4 Method Selection and Relevance

In our Twitter Sentiment Classification project, we have chosen the following five methods from the lists of Explainable AI (XAI) and Causal Inference techniques. These methods were selected based on their strengths, applications, and relevance to our project's objectives.

4.1 Explainable AI Methods

4.1.1 LIME (Local Interpretable Model-Agnostic Explanations)

LIME is an essential tool for providing interpretability to complex machine learning models by explaining individual predictions. This method will allow us to understand specific model decisions on tweet sentiments, offering transparency and insights into the model's behavior. Given the black-box nature of advanced models like LSTM and BERT, LIME will be particularly useful in making these models more interpretable.

4.1.2 SHAP (SHapley Additive exPlanations)

SHAP values provide a comprehensive indication of feature importance by attributing each feature's contribution to the overall prediction. This method is highly relevant for our project as it helps identify which features (words or phrases) significantly influence the sentiment classification, enhancing the transparency and reliability of our models.

4.1.3 Generalized Additive Models (GAMs)

GAMs offer the flexibility to model non-linear relationships while maintaining interpretability. This makes them suitable for capturing complex patterns in tweet sentiments that may not be evident in linear models. By incorporating GAMs, we can better understand the non-linear effects and interactions within our data, providing deeper insights into sentiment trends.

4.2 Causal Inference Methods

4.2.1 Propensity Score Matching

Propensity Score Matching is ideal for reducing bias in observational studies, helping to establish causal relationships between different events or interventions and sentiment changes. In our project, this method will allow us to analyze how specific events (e.g., product launches, political announcements) influence public sentiment on Twitter, offering valuable causal insights.

4.2.2 Regression Discontinuity

The Regression Discontinuity method is effective for estimating causal effects by assigning a cutoff or threshold above and below which treatment is assigned. This technique is relevant for our Twitter sentiment classification project as it enables us to isolate the impact of specific events or interventions on sentiment. By comparing tweets just above and below a cutoff (e.g., a specific date or event threshold), we can precisely measure changes in sentiment and understand the causal effects of these events on public opinion.

4.3 Relevance to the Project

The selected methods collectively provide a robust framework for understanding and interpreting the sentiment classification model's predictions and the underlying causal relationships within the data. By integrating these Explainable AI techniques, we can ensure the transparency and interpretability of our models, making them more trustworthy and actionable. The causal inference methods will help us uncover the causal effects of various events on sentiment, providing deeper insights and supporting data-driven decision-making.

These methods align with our project's objectives to develop a sentiment analysis tool that is not only accurate but also interpretable and insightful, addressing the challenges posed by the diversity and noise of Twitter data.

5 LIME (Local Interpretable Model-Agnostic Explanations)

5.1 Foundational Principles

LIME approximates a complex machine learning model with a local, interpretable model to explain each individual prediction. The technique creates variations of the original instance to understand how the model's prediction changes with small changes in input features. Using these predictions, it trains a simpler model to understand the features' contribution. LIME focuses on specific observations rather than providing a general explanation for the model's behavior, making it highly interpretable for individual data points.

5.2 Procedure

LIME works by perturbing the input data and observing the resulting changes in predictions. It generates synthetic data by making slight modifications to the input features and then observes how the model's predictions vary with these changes. A simpler, interpretable model is then trained on this synthetic data to approximate the complex model's behavior locally around the original instance.

5.3 Strengths

LIME is model-agnostic and open-source, making it applicable to any machine learning model. It is highly interpretable, providing clear reasons for specific predictions. This technique is flexible, handling various data types including numerical, categorical, and text data. LIME also enhances validation and transparency by revealing the effect of individual predictors, and it can be used to examine performance, fairness, and biases in the model.

5.4 Limitations

The effectiveness of LIME relies on the representativeness of perturbations. If the perturbations are not representative, there is a risk of overfitting to the local perturbations rather than providing meaningful explanations. Additionally, LIME can be computationally intensive, especially when working with large models, which may render the process less scalable. Despite its advantages, LIME is still empirically less robust compared to some more established techniques.

5.5 Suitability for Different Types of Data and Business Problems

LIME is flexible and can handle various types of data, making it suitable for a wide range of business problems. It has three core functionalities: the image explainer interprets image classification models, the text explainer provides insights into text-based models, and the tabular explainer assesses the importance of features in tabular datasets. This versatility makes LIME a valuable tool for interpreting models across different domains.

5.6 Application Examples

Alabi, R.O., Elmusrati, M., Leivo, I. et al. demonstrate how LIME can be used to interpret the results of a model predicting the survival of patients with nasopharyngeal cancer. This application showcases LIME's ability to provide clear, interpretable explanations in a critical healthcare context.

5.7 Recent Advancements

Recent advancements in LIME include adaptive perturbation strategies, which refine the perturbation process by tailoring it to the specific characteristics of the data. Additionally, hybrid methods combining LIME with other interpretability techniques, such as SHAP (SHapley Additive exPlanations), have been developed to provide more comprehensive explanations by leveraging the strengths of multiple approaches.

5.8 Conclusion

LIME is a powerful tool for providing local, interpretable explanations for complex machine learning models. Its flexibility and applicability to various data types make it an essential technique for enhancing model transparency and trustworthiness.

6 SHAP (SHapley Additive exPlanations)

6.1 Foundational Principles

SHAP is a technique used to understand feature importance and contributions to predictions, grounded in cooperative game theory concepts where features are considered as players. The Shapley value for a feature represents its average contribution to the model's prediction across all possible combinations of features. SHAP assumes that the model's prediction can be represented as the sum of the contributions of each feature, adhering to an additive model.

6.2 Strengths

SHAP is model-agnostic, meaning it can be applied to any machine learning model. It provides both global insights and local explanations, offering a comprehensive understanding of feature importance. Its theoretical basis ensures the consistency of explanations, leading to accurate representation of feature importance. SHAP also unifies various feature importance measures, making it easier to compare different models and features.

6.3 Limitations

SHAP is computationally complex, especially for models with many features, requiring 2 to the nth power of evaluations. While approximations can reduce complexity, they may lead to measurement errors. The assumption that the model's prediction function is additive might oversimplify the structure of complex models. Additionally, SHAP is empirically less robust compared to some more established techniques.

6.4 Suitability for Different Types of Data and Business Problems

SHAP is highly flexible, capable of handling numerical, categorical, text, and image data. It is suitable for various business problems where understanding feature importance and model transparency is crucial. This versatility makes SHAP a valuable tool across multiple domains.

6.5 Application Examples

- **Credit Scoring**: Explaining the impact of different financial indicators on credit scores.
- **Autonomous Vehicles**: Providing explanations for predictions in autonomous vehicle systems.
- **Healthcare**: Analyzing the influence of different medical parameters on patient outcomes.
- **Interpretable AI for Bio-Medical Applications**: Anoop Sathyan, Abraham Itzhak Weinberg, and Kelly Cohen (2022) used SHAP to explain predictions made by a trained deep neural network to identify benign/malignant masses in breast cancer data. The paper confirmed empirical commonalities in this data for SHAP and LIME techniques.

6.6 Recent Advancements

Recent advancements in SHAP include the development of faster algorithms for calculating Shapley values, which reduce computational complexity. There has been enhanced integration with other machine learning frameworks and tools, streamlining the explanation process. Improved approximation methods have been introduced to handle large datasets and complex models more efficiently.

6.7 Conclusion

SHAP is a powerful tool for providing both global and local explanations of machine learning models. Its flexibility and applicability to various data types and business problems make it an essential technique for enhancing model transparency and understanding feature importance.

7 Generalized Additive Models (GAMs)

7.1 Foundational Principles

Generalized Additive Models (GAMs) extend Generalized Linear Models (GLMs) by allowing more flexibility in modeling non-linear relationships between predictor variables and the response variable. They are additive because each predictor's functional contributions are summed. The specification is $g(E[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$, where each $f_j(X_j)$ is a smooth function. Smoothing functions are often estimated via techniques like splines (e.g., cubic splines, thin-plate splines) or local regression.

7.2 Extensions

- **GAMLSS**: Extends GAMs to include other parameters such as variance, skewness, and kurtosis.
- **Additive Mixed Models**: Incorporate random effects into GAMs to account for hierarchical or grouped data structures.

7.3 Strengths

- **Modeling Flexibility**: GAMs provide a better fit by effectively dealing with nonlinear relationships.
- **Interpretability**: Each predictor's contribution can be examined separately.
- **Reduction of Collinearity**: Each predictor term is modeled separately, reducing collinearity.

7.4 Limitations

- **Additive Restriction**: The model does not inherently include interactions between variables (though they can be added manually).
- **Overfitting Risk**: Despite regularization techniques, careful selection of parameters is necessary to prevent overfitting.
- **Computational Intensity**: GAMs can be computationally intensive.

7.5 Suitability for Different Types of Data and Business Problems

GAMs are suitable for various types of data, including continuous, count, categorical, and spatial data. They are particularly effective in modeling phenomena such as climate change effects (temporal and spatial), risk factor analysis, consumer behavior, sensor data, and experimental data.

7.6 Application Examples

- **Medicine/Finance/Economics**: GAMs are widely used in fields where data can be extremely nonlinear.
- **Environmental Science**: Juliana B. Souza et al. (2018) used a hybrid GAM–PCA–VAR model to quantify the association between respiratory disease and air pollution concentrations.
- **Fishery Science**: Murase et al. (2009) used GAMs to analyze fishery-survey data and reveal the influences of environmental factors on the distribution patterns of Japanese anchovy, sand lance, and krill.

7.7 Recent Advancements

- **Neural Additive Models (NAMs)**: Extend the capabilities of GAMs by integrating neural network structures to capture complex data patterns, addressing limitations like skewness and heteroscedasticity.

7.8 Conclusion

GAMs are powerful tools for modeling non-linear relationships in data, offering flexibility and interpretability. Their applicability to various data types and business problems makes them valuable in enhancing model accuracy and insights.

8 Propensity Score Matching

8.1 Foundational Principles

Propensity Score Matching (PSM) estimates causal effects from observational data by reducing the impact of confounding variables—those associated with both the intervention and outcome. It balances baseline characteristics between treated and untreated subjects, maximizing the likelihood of comparable treatment groups. While randomized controlled trials are ideal, they can be resource-intensive; PSM offers a practical alternative.

8.2 Methodology

PSM calculates propensity scores using a binary model, representing the probability of receiving treatment based on observed characteristics. Subjects are then matched based on these scores using methods like kernel, nearest neighbor, stratification, or radius matching. Key assumptions include partial equilibrium (treatment does not affect control observations), conditional independence (treatment is exogenous), and overlap in characteristics between treated and control groups.

8.3 Strengths

- **Alternative to Randomization**: PSM balances covariates between treated and control groups, reducing selection bias without the cost of randomization.
- **Statistical Power and Flexibility**: Offers greater statistical power than multivariable regression in observational studies with few events and is highly flexible and scalable.
- **No Instrumental Variables Needed**: PSM does not require instrumental variables.

8.4 Limitations

- **Dataset Requirements**: Performs worse in smaller datasets and requires sufficient overlap in propensity scores.
- **Observed Variables Only**: Only accounts for known variables, potentially leaving residual bias. Overfitting can occur with many predictors.
- **Computational Complexity**: Can be computationally intensive, especially with large datasets or complex algorithms.
- **Loss of Data**: Can lead to exclusion of units without suitable matches.

8.5 Suitability for Different Types of Data and Business Problems

PSM is versatile and suitable for use in social sciences, healthcare, public policy, and education. It is highly scalable and best suited for medium to large datasets that include a wide range of covariates.

8.6 Application Examples

- **Health/Psychology/Behavioral Research**: Assessing the effect of teenage alcohol use on educational attainment (Staff, Patrick, Loken, & Maggs, 2008).
- **Education**: Examining the effects of small school size on mathematics achievement (Wyse, Keesler, & Schneider, 2008).

8.7 Recent Advancements

Recent advancements in PSM include achieving covariate balance using machine learning techniques. Ensemble methods like bagged CART, random forests, and boosted CART propensity score models have shown excellent performance in terms of covariate balance and effect estimation, serving as alternatives to logistic regression for estimating propensity scores.

8.8 Conclusion

PSM is a powerful tool for estimating causal effects in observational studies, offering an alternative to randomization. Its flexibility, scalability, and ability to reduce selection bias make it valuable in various fields, from healthcare to public policy.

9 Regression Discontinuity

9.1 Foundational Principles

Regression Discontinuity (RD) is used to estimate causal effects from observational data by assigning a cutoff or threshold above and below which treatment is assigned. Subjects just above the cutoff receive the treatment, while those just below do not. The similarity of observations in all respects except the cutoff seeks to isolate the treatment effect.

9.2 Methodology

The RD methodology involves estimating the model and adding a dummy variable to indicate whether an observation is above the threshold. There are two types of RD designs:

- **Sharp RD**: Deterministic treatment assignment based on the cutoff.
- **Fuzzy RD**: Probabilistic treatment assignment, where the probability of receiving the treatment jumps at the cutoff.

9.3 Strengths

- **Quasi-Experimental Nature**: Establishes credible causal relationships by comparing observations around the cutoff.
- **Modeling Flexibility**: Allows for both parametric and non-parametric approaches without requiring a specific form.
- **Intuitive Design**: Easy to interpret and communicate, with visual validation.

9.4 Limitations

- **Statistical Power**: Requires a large number of observations around the cutoff, with potential for noisy estimates.
- **Manipulation Risk**: Results can be invalid if individuals manipulate the "assignment variable."
- **Local Estimates**: RD estimates may be local and not generalizable to other values of the running variable.

9.5 Suitability for Different Types of Data and Business Problems

RD is a versatile technique suitable for various data types where the key assumption is that the relationship between the running variable and the outcome is continuous around the cutoff. It is crucial to choose the range of data carefully, as the treatment effect is often assumed to be homogeneous around the cutoff. RD can be applied to business problems such as loan approval, policy changes, scholarships, and regulatory policy.

9.6 Application Examples

- **Public Policy/Healthcare/Education/Economics**: Hausman and Rapson (2018) assessed the impact of tax, social spending, financial aid, and health outcomes using a temporal RD design to evaluate the impact of changes in energy prices on consumption patterns.

9.7 Recent Advancements

- **Geographic RD Designs**: Keele and Titiunik (2015) used geographic boundaries as cutoffs to investigate the effects of the Voting Rights Act.
- **Integration with Machine Learning**: Knaus, Lechner, and Strittmatter (2020) integrated random forests and gradient boosting on labor data.
- **Manipulation Testing**: Improved techniques for testing manipulation of the running variable have been developed, including detection of discontinuities in the density of the running variable.

9.8 Conclusion

RD is a powerful tool for estimating causal effects in observational studies, offering a quasi-experimental approach. Its intuitive design, flexibility, and applicability to various fields make it a valuable method for understanding the impact of treatments and policies.

References

[1] Neethu, M.S., Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *Proceedings of 4th ICCCNT*, 1-5. https://ieeexplore.ieee.org/document/6726818

[2] Basiri, M.E., Nemati, S., Abdar, M., Cambria, E., Acharya, U.R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems*, 115, 279-294. https://www.sciencedirect.com/science/article/abs/pii/S0167739X20309195

[3] Krugmann, J.O., Hartmann, J. (2024). Sentiment Analysis in the Age of Generative AI. *Customer Needs and Solutions*, 11, 3. https://link.springer.com/article/10.1007/s40547-024-00143-4

[4] Lee, B.K., Lessler, J., Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29, 337–346. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/

[5] Staff, J., Patrick, M.E., Loken, E., Maggs, J.L. (2008). Teenage alcohol use and educational attainment. *Journal of Studies on Alcohol and Drugs*, 69, 848–858. https://pubmed.ncbi.nlm.nih.gov/18925343/

[6] Lalani, N., Jimenez, R.B., Yeap, B. (2020). Understanding Propensity Score Analyses. *Int J Radiat Oncol Biol Phys*, 107(3), 404-407. doi: 10.1016/j.ijrobp.2020.02.638. PMID: 32531385. https://www.redjournal.org/article/S0360-3016(20)30888-9/pdf

[7] Wyse, A.E., Keesler, V., Schneider, B. (2008). Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. *Teachers College Record*, 110, 1879–1900. https://www.researchgate.net/publication/269337014_Assessing_the_effect_of_small_school_size_on_mathematics_achievement_A_propensity_score_approach

[8] Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55. https://www.stat.cmu.edu/~ryantibs/journalclub/rosenbaum_1983.pdf

[9] Lee, D.S., Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2), 281–355. doi: 10.1257/jel.48.2.281. https://www.aeaweb.org/articles?id=10.1257/jel.48.2.281

[10] Angrist, J.D., Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Chapter 6. https://www.researchgate.net/publication/51992844_Mostly_Harmless_Econometrics_An_Empiricist's_Companion

[11] Hausman, C., Rapson, D. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10, 533-552. https://www.annualreviews.org/content/journals/10.1146/annurev-resource-121517-033306

[12] Keele, L., Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1), 127-155.

[13] Knaus, M.C., Lechner, M., Strittmatter, A. (2020). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 23(2), 76-91.

[14] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[15] Tutz, G., Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics*, 62, 961–971.

[16] Thielmann, A., Kruse, R.M., Kneib, T., Säfken, B. (2023). Neural Additive Models for Location Scale and Shape: A Framework for Interpretable Neural Regression Beyond the Mean. https://ar5iv.labs.arxiv.org/html/2301.11862

[17] Murase, H., Nagashima, H., Yonezaki, S., Matsukura, R., Kitakado, T. (2009). Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: a case study in Sendai Bay, Japan. *ICES Journal of Marine Science*, 66, 1417–1424.

[18] Souza, J.B., Reisen, V.A., Franco, G.C., Ispány, M., Bondon, P., Santos, J.M. (2018). Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67(2), 453–480. https://academic.oup.com/jrsssc/article/67/2/453/7058313

[19] Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

[20] Dieber, J., Kirrane, S. (2020). Why Model Why? Assessing the Strengths and Limitations of LIME. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT 2020)*. Association for Computing Machinery. https://arxiv.org/pdf/2012.00093

[21] Alabi, R.O., Elmusrati, M., Leivo, I. et al. (2023). Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Scientific Reports*, 13, 8984. https://doi.org/10.1038/s41598-023-35795-0

[22] Li, X., Bing, L., Lam, W., Shi, B. (2019). Transformation Networks for Target-Oriented Sentiment Classification. *ArXiv*. https://arxiv.org/abs/1805.01086

[23] Lundberg, S.M., Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions https://arxiv.org/abs/1705.07874