Twitter Sentiment Classification

Alen Pavlovic, Stepan Ochodek, Esteban Felix Tapia, Jeff Clancy

Masters of Applied Data Science
University of Chicago

August, 2024

Background & Problem Statement

Background and Context

- Social media platforms like Twitter have become rich sources of real-time information, capturing the diversity of human expression across various contexts. This project leverages the Sentiment140 dataset, a large collection of tweets annotated with sentiment labels, to develop a sentiment analysis tool.

Problem Statement

- The primary challenge is to accurately classify the sentiment of tweets (positive or negative) using a combination of linear and non-linear machine learning models. This classification must be effective even for texts that do not contain explicit emoticons.

Objectives

1. **Develop a Sentiment Analysis Tool**
   Create a tool that can classify sentiments in tweets using both traditional and advanced machine learning models.
2. **XAI Techniques**
   Use XAI methods like LIME and SHAP to make the models decision-making process transparent and interpretable.
3. **Evaluate Model Performance**
   Use cross-validation and independent test sets to assess the performance of the models.
4. **Causal Analysis**
   Apply causal inference techniques to uncover and understand causal relationships within the data.

1. Introduction
2. **Literature Review**
3. Methods
4. Results
5. References

Sentiment Analysis in Twitter Using Machine Learning Techniques

Challenge

- Informal language & short tweets present difficulties for classification models

Proposed Solutions

- **Preprocessing:** Removing URLs, handling slang, and correcting misspellings
- **Feature Extraction:** Extraction of Twitter-specific features like hashtags
- **Classification Models post Preprocessing & Feature Extraction:**
  - Naive Bayes (NB)
  - Support Vector Machine (SVM)
  - Maximum Entropy (ME)
  - Ensemble Classifier: Combines NB, SVM, and ME using a voting mechanism

Relevance / Adaptability

- Preprocessing & Feature Extraction are used in the model in this presentation

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

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

Challenge

- Individual sentiment classification techniques can have limitations, especially when translating short form methods to long form text

Proposed Solutions

- **CNNs and RNNs:** Identify word & phrase 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.

Relevance / Adaptability

- Provides a set of advanced techniques that can be employed if effectiveness of initial approach is limited or if model is scaled to be effective on long form text

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

Sentiment Analysis in the Age of Generative AI

Challenge

- Initial sentiment classification models may have limitations from being trained in a specific, closed environment; LLMs may not have this limitation

Proposed Solutions

- Compare classification from GPT-3.5, GPT-4, and Llama 2 against traditional transfer learning models like SiEBERT and fine-tuned RoBERTa
- The LLMs were able to perform similarly to SiEBERT and fine-tuned RoBERTa, showing that they are a viable alternative for sentiment classification

Relevance / Adaptability

- GPT-4 and Llama 2 were considered as a supplement to the project in case there was a need to simplify the process of adapting the model to new types of text data by leveraging the pre-trained capabilities of these models.
- The paper stressed the importance of preprocessing to handle text complexity and structured content

Source: 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.

Title

- Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante.

Microsoft® Windows	Apple® Mac OS
Windows-Kernel	Unix-like
Arm, Intel	Intel, Apple Silicon
Sudden update	Stable update
Less security	More security
...	...

Algorithms

Non-Numbering Formula

$$J(\theta) = \mathbb{E}_{\pi_\theta}[G_t] = \sum_{s \in \mathcal{S}} d^\pi(s) V^\pi(s) = \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) Q^\pi(s, a)$$

Multi-Row Formula[1]

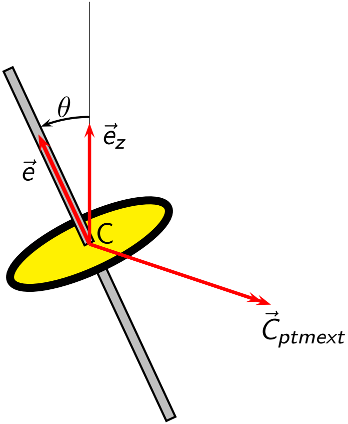
$$\begin{aligned} Q_{\mathrm{target}} &= r + \gamma Q^\pi(s', \pi_\theta(s') + \epsilon) \\ \epsilon &\sim \mathrm{clip}(\mathcal{N}(0, \sigma), -c, c) \end{aligned} \tag{1}$$

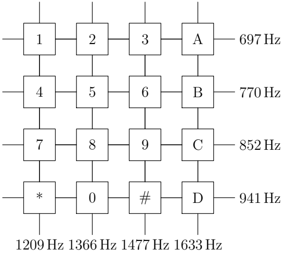
[1] If text appears in the formulause \mathrm{} or \text{} instead

Numbered Multi-line Formula

$$
\begin{multline}
A = \lim_{n\rightarrow\infty} \Delta x \left(a^2 + \left(a^2 + 2a\Delta x + (\Delta x)^2 \right) \right. \\
+ \left(a^2 + 2\cdot 2a\Delta x + 2^2 (\Delta x)^2 \right) \\
+ \left(a^2 + 2\cdot 3a\Delta x + 3^2 (\Delta x)^2 \right) \\
+ \ldots \\
\left. + \left(a^2 + 2\cdot (n-1)a\Delta x + (n-1)^2 (\Delta x)^2 \right) \right) \\
= \frac{1}{3} \left(b^3 - a^3 \right) \tag{2}
\end{multline}
$$

Graphics and Columns





LaTeX Common Commands

Commands

\chapter chapter	\section section	\subsection sub-section	\paragraph paragraph
\centering center	\emph emphasize	\verb original	\url hyperlink
\footnote footnote	\item list item	\caption caption	\includegraphics insert image
\label label	\cite citation	\ref refer	

Environment

table table	figure figure	equation formula
itemize non-numbering item	enumerate numbering item	description description

LaTeX Examples of environmental commands

```latex
\begin{itemize}
  \item A \item B
  \item C
  \begin{itemize}
    \item C-1
  \end{itemize}
\end{itemize}
```

- A
- B
- C
  - C-1

LaTeX Examples of environmental commands

```latex
\begin{itemize}
  \item A \item B
  \item C
  \begin{itemize}
    \item C-1
  \end{itemize}
\end{itemize}
```

- A
- B
- C
  - C-1

```latex
\begin{enumerate}
  \item A \item B
  \item C
  \begin{itemize}
    \item[n+e]
  \end{itemize}
\end{enumerate}
```

1. A
2. B
3. C
   - n+e

LaTeX Formulas

```latex
$V = \frac{4}{3}\pi r^3$

\[
  V = \frac{4}{3}\pi r^3
\]

\begin{equation}
  \label{eq:vsphere}
  V = \frac{4}{3}\pi r^3
\end{equation}
```

$V = \frac{4}{3}\pi r^3$

$$V = \frac{4}{3}\pi r^3$$

$$V = \frac{4}{3}\pi r^3 \tag{3}$$

- more information here

```latex
\begin{table}[htbp]
  \caption{numbers & meaning}
  \label{tab:number}
  \centering
  \begin{tabular}{cl}
    \toprule
    number & meaning \\
    \midrule
    1 & 4.0 \\
    2 & 3.7 \\
    \bottomrule
  \end{tabular}
\end{table}
```

Table 1: numbers & meaning

numbers	meaning
1	4.0
2	3.7

formula (3) at previous slide and Table 1

1. Introduction
2. Literature Review
3. Methods
4. **Results**
5. References

- Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit.
- In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat.
- Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam.
- Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi.

[1] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," in *Stanford*, 2009.

[2] TensorFlow Datasets, "Sentiment140."

[3] M. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *Proceedings of 4th ICCCNT*, pp. 1–5, 2013.

[4] M. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. Acharya, "ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Generation Computer Systems*, vol. 115, pp. 279–294, 2021.

[5] J. Krugmann and J. Hartmann, "Sentiment analysis in the age of generative ai," *Customer Needs and Solutions*, vol. 11, no. 3, 2024.

[6] B. Lee, J. Lessler, and E. Stuart, "Improving propensity score weighting using machine learning," *Statistics in Medicine*, vol. 29, pp. 337–346, 2010.

[7] J. Staff, M. Patrick, E. Loken, and J. Maggs, "Teenage alcohol use and educational attainment," *Journal of Studies on Alcohol and Drugs*, vol. 69, pp. 848–858, 2008.

[8] N. Lalani, R. Jimenez, and B. Yeap, "Understanding propensity score analyses," *International Journal of Radiation Oncology Biology Physics*, vol. 107, no. 3, pp. 404–407, 2020.

[9] A. Wyse, V. Keesler, and B. Schneider, "Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach," *Teachers College Record*, vol. 110, pp. 1879–1900, 2008.

[10] P. Rosenbaum and D. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, pp. 41–55, 1983.

[11] D. Lee and T. Lemieux, "Regression discontinuity designs in economics," *Journal of Economic Literature*, vol. 48, no. 2, pp. 281–355, 2010.

[12] J. Angrist and J. Pischke, *Mostly Harmless Econometrics: An Empiricists Companion*.
Princeton University Press, 2009.

[13] C. Hausman and D. Rapson, "Regression discontinuity in time: Considerations for empirical applications," *Annual Review of Resource Economics*, vol. 10, pp. 533–552, 2018.

[14] L. Keele and R. Titiunik, "Geographic boundaries as regression discontinuities," *Political Analysis*, vol. 23, no. 1, pp. 127–155, 2015.

[15] M. Knaus, M. Lechner, and A. Strittmatter, "Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence," *The Econometrics Journal*, vol. 23, no. 2, pp. 76–91, 2020.

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

[17] G. Tutz and H. Binder, "Generalized additive modelling with implicit variable selection by likelihood based boosting," *Biometrics*, vol. 62, pp. 961–971, 2006.

[18] A. Thielmann, R. Kruse, T. Kneib, and B. Säfken, "Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean," 2023.

[19] H. Murase, H. Nagashima, S. Yonezaki, R. Matsukura, and T. Kitakado, "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*, vol. 66, pp. 1417–1424, 2009.

[20] J. Souza, V. Reisen, G. Franco, M. Ispány, P. Bondon, and J. Santos, "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*, vol. 67, no. 2, pp. 453–480, 2018.

[21] M. Ribeiro, S. Singh, and C. Guestrin, "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, 2016.

[22] J. Dieber and S. Kirrane, "Why model why? assessing the strengths and limitations of LIME," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT 2020)*, 2020.

[23] R. Alabi, M. Elmusrati, I. Leivo, and et al., "Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP," *Scientific Reports*, vol. 13, p. 8984, 2023.

[24] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," *ArXiv*, 2019.

[25] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *ArXiv*, 2017.

Thank You