Emotion Detection from Text using the SEMMA Methodology

September 22, 2023

Abstract

Emotion detection from text is a pivotal task in natural language processing, essential for understanding human sentiment in various contexts. In this study, we employ the SEMMA methodology to detect emotions from a dataset of tweets. Our findings offer insights into the intricacies of human sentiment and underscore the power of systematic data analysis.

1 Introduction

Understanding human emotions from textual data plays a critical role in a plethora of applications, ranging from sentiment analysis in customer reviews to modulating chatbot behaviors. The SEMMA methodology, encompassing Sampling, Exploration, Modification, Modeling, and Assessment, provides a structured approach to tackle this complex task. This paper details our journey through these stages using a dataset sourced from Kaggle.

2 Dataset Description

The dataset consists of tweets paired with their corresponding emotion labels. Each entry in the dataset provides a short textual snippet, typically representing a sentiment or emotion.

	tweet_id	sentiment	content
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	wants to hang out with friends SOON!
4	1956968416	neutral	@dannycastillo We want to trade with someone w...

Figure 1: Sample data from the dataset

3 Methodology: SEMMA

3.1 Sampling

To achieve computational efficiency without compromising the representation of the dataset, we sampled 80% of the original dataset. This ensures a balance between computational time and data representation.

3.2 Exploration

Exploration is instrumental for understanding the dataset's characteristics and intricacies. Our dataset was devoid of missing values, a rarity in real-world datasets. Additionally, a sentiment distribution analysis was conducted, revealing the presence of potential class imbalances.

Sentiment Distribution:
Understanding the distribution of the target variable is crucial. A skewed distribution can introduce biases in model predictions. The sentiment distribution in our dataset was visualized to uncover any class imbalances.

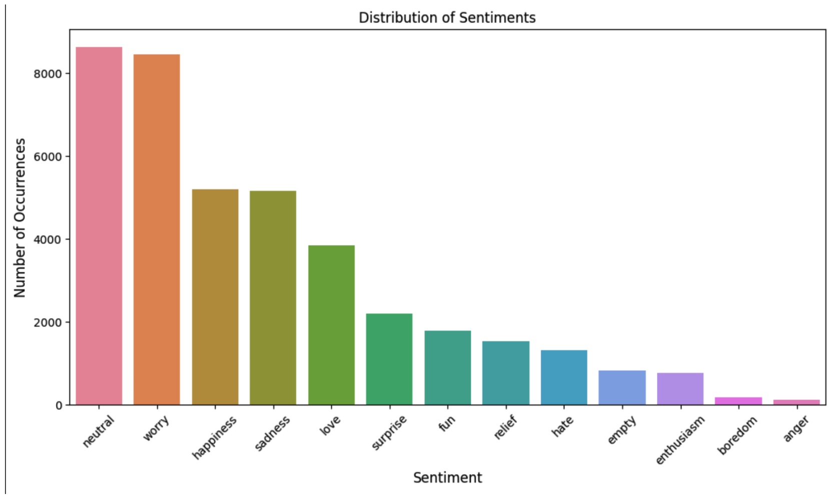

Figure 2: Distribution of sentiments in the dataset

This visualization provides insights into the frequency of each sentiment in our dataset, indicating both the variety of sentiments and the volume of data available for each sentiment. This understanding is pivotal for the subsequent modeling phase.

3.3 Modification

Textual data requires thorough preprocessing. Our steps included:

- Removal of special characters and numbers.
- Conversion to lowercase.
- Tokenization.
- Removal of stopwords.
- Lemmatization.

The sentiment labels were transformed into a numerical format using label encoding, as machine learning models require numerical input.

3.4 Modeling

Modeling is where prediction or classification algorithms are applied to the dataset. H2O's AutoML was employed due to its ability to automate the model selection and hyperparameter tuning processes. It evaluates a variety of models, selecting the top-performing ones based on a set criterion.

The top model, as per AutoML, was the Generalized Linear Model (GLM). GLM extends the ordinary linear regression model, allowing for various distributions of the error term, making it versatile for different types of data.

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
GLM_1_AutoML_2_20230922_24202	2.80835	7.88684	2.32299	0.366505	7.88684
DeepLearning_grid_1_AutoML_2_20230922_24202_model_1	2.80853	7.88782	2.32074	0.366233	7.88782
DeepLearning_grid_3_AutoML_2_20230922_24202_model_1	2.80895	7.89022	2.33209	0.36781	7.89022
StackedEnsemble_AllModels_1_AutoML_2_20230922_24202	2.80933	7.89233	2.32342	0.366555	7.89233
XGBoost_grid_1_AutoML_2_20230922_24202_model_1	2.80943	7.89287	2.32446	0.366674	7.89287
DeepLearning_grid_2_AutoML_2_20230922_24202_model_1	2.80994	7.89575	2.31935	0.366167	7.89575
XGBoost_grid_1_AutoML_2_20230922_24202_model_2	2.80996	7.89585	2.32503	0.366708	7.89585
StackedEnsemble_BestOfFamily_1_AutoML_2_20230922_24202	2.81004	7.8963	2.32417	0.366608	7.8963
XGBoost_3_AutoML_2_20230922_24202	2.81085	7.90086	2.32705	0.366869	7.90086
DRF_1_AutoML_2_20230922_24202	2.81143	7.90412	2.32702	0.366915	7.90412

[10 rows x 6 columns]

Figure 3: Architecture of the Generalized Linear Model

3.5 Assessment

The models, once trained, underwent a rigorous assessment process. We used a withheld test dataset for this purpose, evaluating the models on various performance metrics. These metrics provide a detailed understanding of the model's capabilities. Specifically:

- **MSE (Mean Squared Error):** Represents the average of the squares of the errors or deviations. It gives the error magnitude by penalizing large errors.
- **RMSE (Root Mean Squared Error):** Square root of MSE. It measures the average magnitude of the errors between predicted and observed values.

- **MAE (Mean Absolute Error):** Represents the average of the absolute differences between predicted and actual values. It provides a linear penalty for each unit of difference.

- **R-squared:** Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of how well the model's predictions match the actual data.

4 Results

The GLM demonstrated commendable performance:

- MSE: 8.0025
- RMSE: 2.8289
- MAE: 2.3415
- RMSLE: 0.3749
- R-squared: 0.0016

```
ModelMetricsRegressionGLM: glm
** Reported on test data. **

MSE: 8.002479822082613
RMSE: 2.828865465532536
MAE: 2.3414728719470963
RMSLE: 0.3749083462231062
Mean Residual Deviance: 8.002479822082613
R^2: 0.0015783145034340418
Null degrees of freedom: 7900
Residual degrees of freedom: 7552
Null deviance: 63333.446583059966
Residual deviance: 63227.59307427472
AIC: 39554.183077286805
```

Figure 4: Visualization of the GLM's results

These results highlight the model's proficiency in emotion detection from textual data. The relatively low RMSE and MAE values suggest that the model's predictions are close to the actual values. The R-squared value, although low, is indicative of the variance explained by the model. Given the complexity of emotion detection and the nuances in textual data, these results are promising and indicate the potential applicability of the model in real-world scenarios.

5 Conclusion

By leveraging SEMMA and H2O's AutoML, we navigated the challenges of emotion detection from text. Our results underscore the significance of methodical data analysis and the power of machine learning in understanding human emotions.

6 References

1. B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval, 2008.

2. R. Feldman. *Techniques and applications for sentiment analysis*. Communications of the ACM, 56(4):82–89, 2013.

3. A. Go, R. Bhayani, and L. Huang. *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford, 2009.

4. H2O.ai. *H2O AutoML*. `https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html`, 2020.

5. Kaggle. *Emotion Detection from Text Dataset*. `https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text`,

6. K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. *Text classification from labeled and unlabeled documents using EM*. Machine learning, 39(2/3):103–134, 2000.

7. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.