Computer Classification of News

Siyi Liu, Supervised by Dr. George Tzanetakis
University of Victoria, Department of Computer Science, March 2024

This research was supported by the
Jamie Cassels Undergraduate Research Award

INTRODUCTION

In 2019, 970 million people around the world were living with a mental disorder[1]. Psychological counseling services are very expensive, especially for underdeveloped countries and regions. If we want to create robots that can provide psychological services for humans, the first step is to enable robots to understand human emotions. Therefore, my JCURA research goal is to enable computers to read text and determine the emotions expressed in that text. The dataset was chosen randomly to test if my classification works.

DATA SET

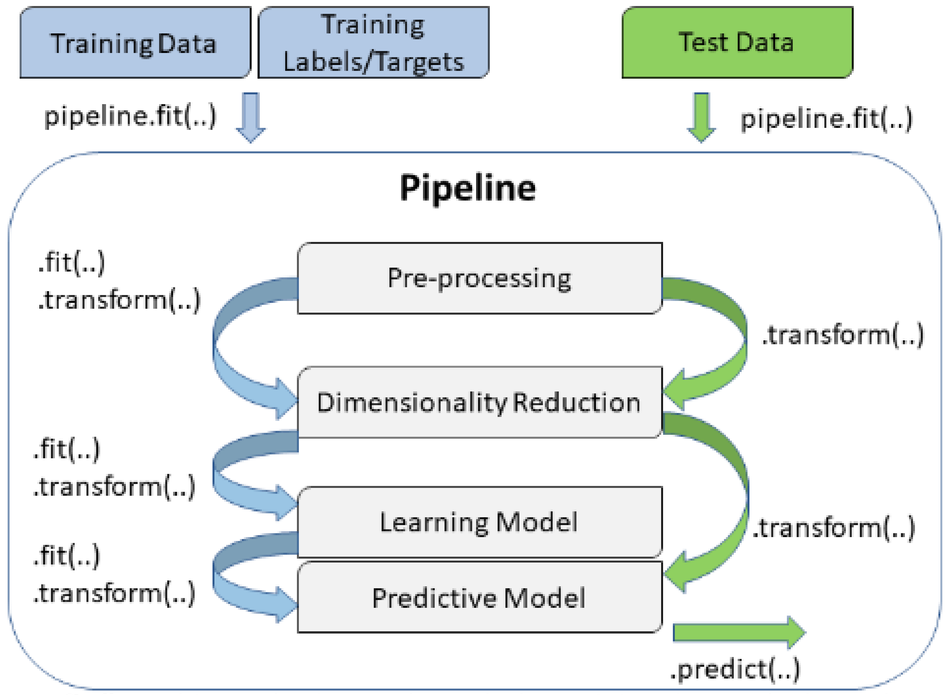
The AG News(Tags: “World”, “Sports”, “Business”, “Sci/Tech”) contains 30,000 training and 1,900 test samples per class. [2]

lable	text
1 World	France marks the 'other D-Day' Two days of celebrations to honour the Allied veterans who liberated southern France near a climax.
2 Sports	US NBA players become the Nightmare Team after epic loss (AFP) AFP - Call them the "Nightmare Team".
3 Business	Wall St. Bears Claw Back Into the Black (Reuters) Reuters - Short-sellers, Wall Street's dwindling\band of ultra-cynics, are seeing green again.
4 Sci/Tech	Oracle expands midmarket ambitions Company looks to juice its application server business with a version tuned for smaller organizations.

METHODS

i. Remove common words, such as articles, using a Python library.

ii. Scikit-learn Pipelines API [3]



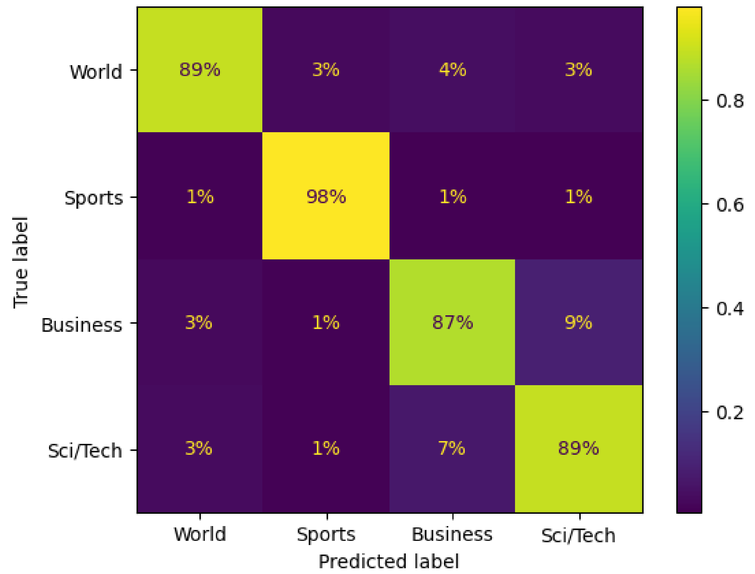
- **fit ()** — This method goes through the training data, calculates the parameters (like mean (μ) and standard deviation (σ) in StandardScaler class) and saves them as internal objects.
- **transform()** — The parameters generated using the fit() method are now used and applied to the training data to update them.

RESULT USING SVM CLASSIFIER

	precision	recall	f1-score	support
World	0.94	0.91	0.92	1900
Sports	0.96	0.99	0.97	1900
Business	0.90	0.89	0.89	1900
Sci/Tech	0.90	0.91	0.90	1900
accuracy			0.92	7600
macro avg	0.92	0.92	0.92	7600
weighted avg	0.92	0.92	0.92	7600

CONFUSION MATRIX

Confusion Matrixes is about what types of errors the classifier makes



precision=TP/(TP+FP);
recall(TPR)=TP/(TP+FN) ;
Accuracy=(TP+TN)/(TP+TN+FP+FN)
F1 measure= 2/(1/precision+1/recall)
=2×(precision×recall))/(precision+recall)

CONCLUSIONS

Test set contains 1900 data, and the accuracy of prediction using SVM classifier reached 92%. The AI tool can predict human emotions based on the interpretation of the text.

REFERENCES

[1] “Mental disorders,” World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders. (Accessed: 05-Mar-2024)
[2] “Papers with Code - AG News Dataset,” AG News Dataset | Papers With Code. [Online]. Available: https://paperswithcode.com/dataset/ag-news. (Accessed: 05-Mar-2024)
[3] Rachidyz, “Tutorial scikit-learn || zero to hero || part I,” Kaggle, 23-Nov-2020. [Online]. Available: https://www.kaggle.com/code/rachidyz/tutorial-scikit-learn-zero-to-hero-part-i. (Accessed: 09-Mar-2024)

ACKNOWLEDGEMENTS

- I am deeply grateful to my supervisor Dr. George Tzanetakis for his willingness, and patience in guiding me through this research project, as well as for imparting valuable insights into how to conduct research.
- Also, I am immensely grateful to Nancy Ami from the Centre for Academic Communication for her tremendous help and support this year. I would also like to express my sincere appreciation to Nancy for assisting me in planning and creating this poster.
- Lastly, I would like to thank the Learning and Teaching Support Innovation office for providing me this JCURA