News Authentication Via Emotion (NAVE)

Alexander Epstein
Northeastern University
Boston, MA 02115
alexanderepstein@tuta.io

Abstract

We developed a method for utilizing a large language model (LLM) as a way to generate intermediate interpretable features for the downstream task of fake or real news classification. The LLM generates fine grained emotion sentiment analysis on both the title and each sentence of the article from which the features are exracted. Then these features are processed through a simpler classifier to determine the authenticity of the sample news article. Results of the final classifier indicate a test accuracy of ~89% suggesting the practical usage of this model in authenticating news articles.

1 Introduction

Machine Learning advancements have enabled large performance improvements in various problems spanning both scientific research and business. However one of the pitfalls of complex architectures that have the ability to perform complex tasks are rarely interpretable and even less so in Natural Language Processing. It is desired to provide insight into the way that these models perform these tasks so that the outputs need not be taken immediately at face value. This interpretability is instrumental in the adoption of Artificial intelligence in the mainstream and ensures we have both fair and ethical intelligent systems.

In the Age of Information it is shown that retractions in the scientific community are detrimental to the societies understanding of the topics as retractions rarely garner the same interest as the original misinformation [1]. It is not a large leap to say that this phenomenon would extend to news as well which can lead to disinformation spreading rapidly across our society. If we seek to stop the spread of this misinformation we need to intervene at the moment of either production or consumption of this information. This work seeks to tackle this problem from the standpoint of consumption as this does not require any integration with the publishers and can be applied more generically.

In this vein we saught to authenticate the trustworthiness of a news article utilizing interpetable features. This has a two-fold impact, it allows us to attempt to provide information to the consumer that grades the trustworthiness of the information possibly preventing the spread of misinformation, while attempting to make the models reasoning more interpretable, allowing for the consumer to decide if they trust the output of the algorithm.

2 Method

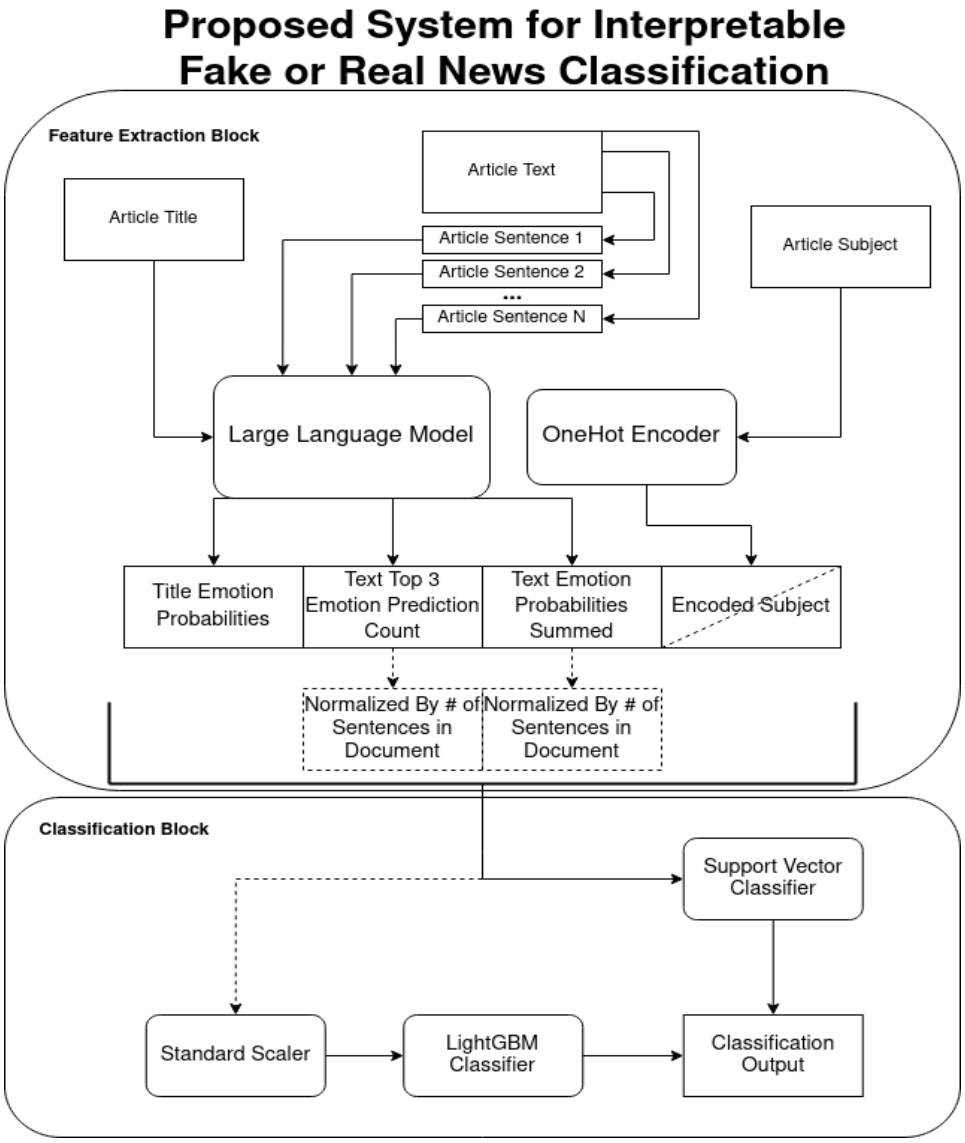

Figure 1: System proposed for interpretable news authenticity classification. Dotted lines represent a variation of the original architecture that is used in the final implementation.

2.1 Emotion Sentiment Analysis

To create our interpretable features we consider fine-grained emotion sentiment analysis as a feature extraction step for our downstream classifier. For this task we fine tune a pretrained Large Language Model on the GoEmotions Dataset[2].

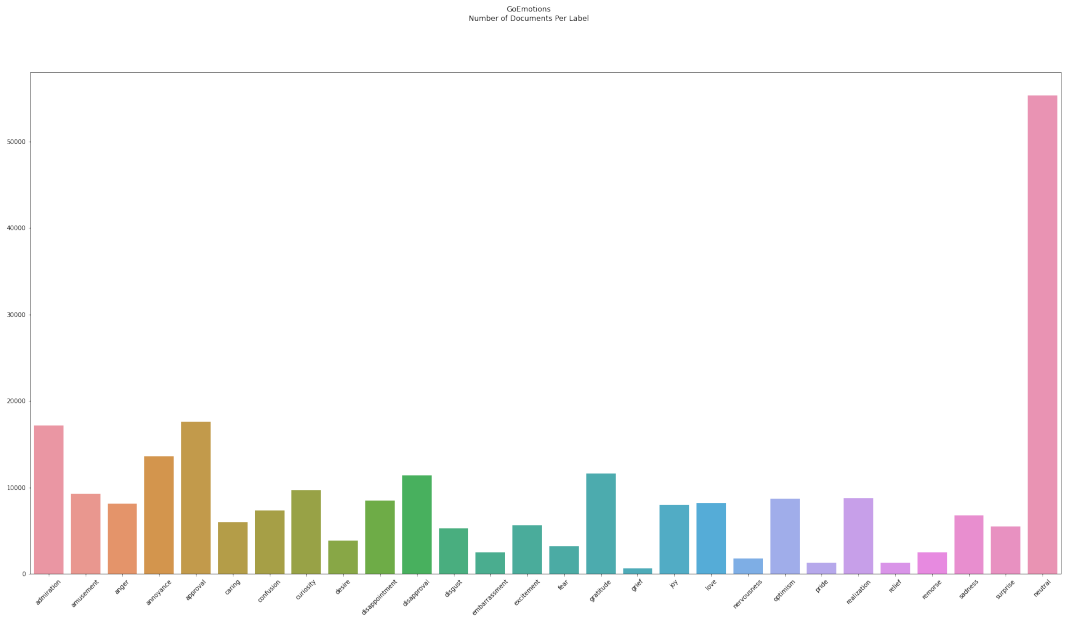

Figure 2: GoEmotions number of documents per label. Preliminary analysis to determine priors of dataset.

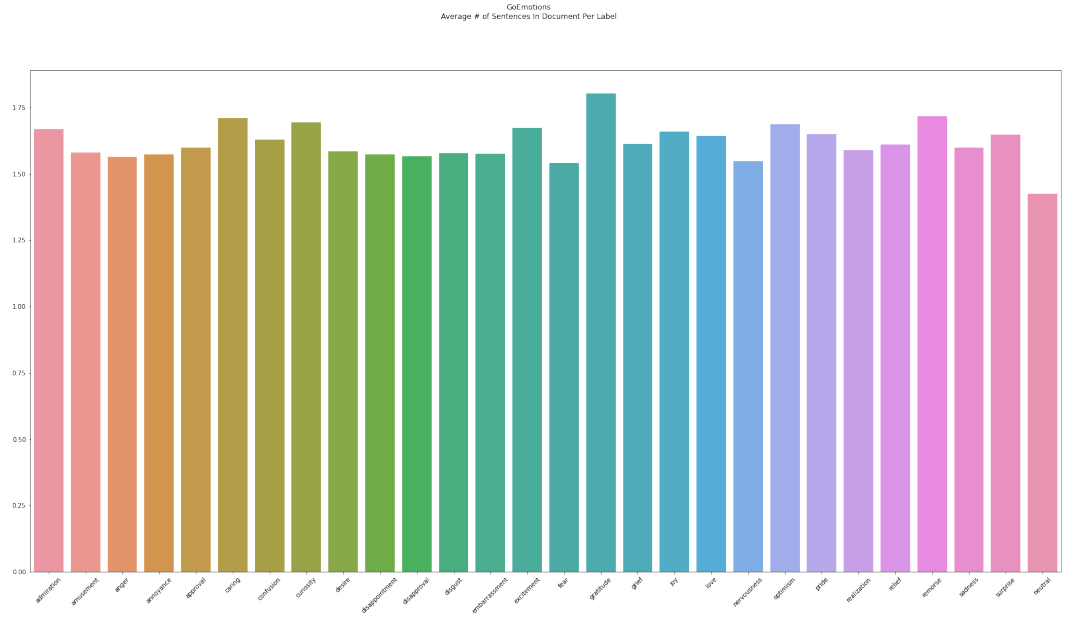

Figure 3: GoEmotions Average # of Sentences in Document Per Label. Preliminary analysis to understand performance bias on downstream emotional sentiment analysis per label.

In Figure 2 we analyze the counts of the emotion labels across the dataset. We can consider this to be a bias in the dataset, however it can also be seen as a prior that is representative of the real distribution from which the labels are sampled. In Figure 3 we look at the average number of sentences per document for documents with the corresponding label. This shows that the model will not be exposed to a bias relating to the length of the text being analyzed.

With these two analysis we decided to not perform any label or sentence count based under/ over sampling techniques and to process the data as is. We utilize the pretrained tokenizer corresponding to our choice of Large Language Model, Cased DistillBERT.

For model fine tuning we perform a 70, 20, 10 for the train, evaluation, test split. We utilize the Adam Optimizer with a Sparse Categorical Cross Entropy loss and train for 5 epochs with a batch size of 128.

2.2 Fake or Real News Classification

To enable our network to classify the authenticity of news articles we saught a dataset that contained hard authenticity labels. The Fake or Real News dataset[3, 4] is comprised of 21417 real and 23481 fake news articles, along with their corresponding title and subject.

As seen in the upper portion of Figure 1 we handle both the title and the article text by processing them through the LLM. The title is processed directly while the article is first tokenized into sentences and we run the model on each sentence independently. From the title we utilize the raw probability outputs. For the article sentences we extract two features the first is the probabilities of each label summed across the sentences. The second feature is we take the count for each label showing up within the top three predictions. We then originally passed the article subject through a one-hot encoder as an extra set of features. However that was shown to introduce a large bias in our dataset. So to remove this bias we removed this from the feature set and we normalized the two extracted features from the text by the number of sentences in that corresponding document.

In the bottom half of Figure 1 we see the two methods for classification that correspond to the two methods utilized for constructing the feature vector. For the feature vector that contains sentence related features unnormalized by the number of sentences along with the encoded subject we utilized a Support Vector Classifer with a radial basis function kernel. For the feature vector that removes the encoded subject and normalizes the sentence related features by the number of sentences we use a pipeline of a Standard Scaler which performs simple mean and variance normalization followed by a LightGBM Classifier.

3 Results & Analysis

3.1 Emotion Sentiment Analysis

The performance of the LLM on the dataset is okay and is similar to the performance that was measured by the authors of the dataset for their model. We can see various classification metrics across the labels in Figure 4. In Figure 7 we see the confusion matrix for the model outputs across the aplit datasets. What is most apparent in the confusion matrix is that neutral is the label most commonly predicted for text with a ground truth of a different label. This is most likely due to the bias of this emotion in our dataset as seen in Figure 2.

While the model performs as expected to make the application more practical in the wild we need to be able to cut down on model inference time. We can attempt to do so by either further distilling this model into a smaller model with a student teacher framework and we can look to do Quantization Aware Training along with pruning. If we could severly cut down on

the model size and inference time then it would be reasonable to have this model run client-side when analyzing the authenticity of an article.

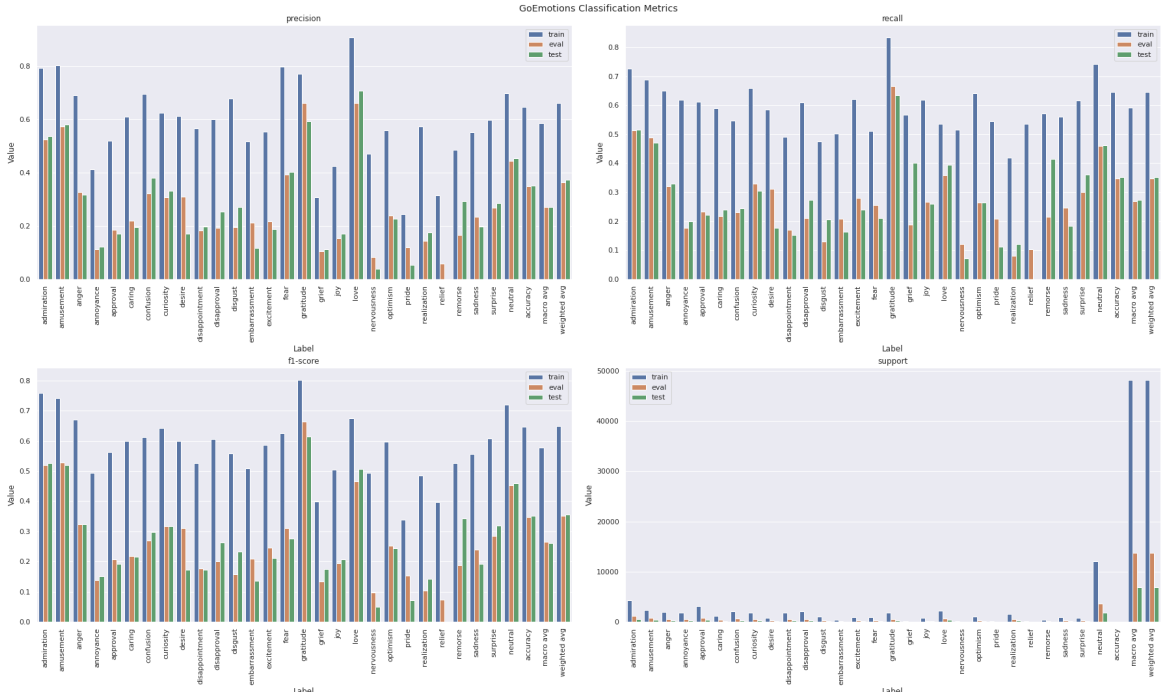

Figure 4: Classification metrics from DistillBERT on GoEmotions Dataset.

3.2 Fake Or Real News

Once we have constructed our set of features according to the architecture in Figure 1 we analyze the TSNE embeddings of these features in Figure 9. Additionally we also plot the correlation between all the features and the label as seen in Figure 10. In these correlation matrices we only need to look at either the upper right or lower left triangle as these are mirror images of one another. While there are many interesting patterns here by looking at the last row of the confusion matrix we see the correlation between the features and the label. We see a very strong correlation when it comes to the encoded subject of the feature set. If we take these features and pass them through a Support Vector Machine (SVM) we get almost perfect results across all three datasets achieving precision, recall and f1 scores of ~99%, the breakdown of these results is in Figure 6 with the confusion matrix in Figure 12. This result is too good and by looking further into the dataset we see that the subjects used for fake and real news are disjoint. In other words just by having the encoded subject feature alone we would be able to achieve these near-perfect results. We can also see this show up in the TNSE embedding present in Figure 9. The points that are disjoint from the cluster most likely derive from the encoded subject.

To remove this bias we look to reconstruct these features to make the results more comparable to the application of the model in the wild that would not have this bias. In doing so we follow the feature extraction path in dotted lines seen in Figure 1. Now when we look at the correlation matrix and TSNE embeddings of these new features in Figure 8 & 11 we see that this bias is removed. We also see that some of the feature colinearity is removed as well which is done by normalizing the text predictions and probabilities by the number of sentences in that respective document. With these new features we first utilized the same

SVM as with the original feature set and our performance dropped by more than 25% across all the data splits. In order to improve this performance after some experimentation we settled on a pipeline of a Standard Scaler along with a LightGBM Classifier and we end up with the classification metrics seen in Figure 5. These results are more reasonable with a near-perfect precision, recall and f1 score on the training set and around 90% for the precision, recall, and f1 scores for the evaluation and testing set.

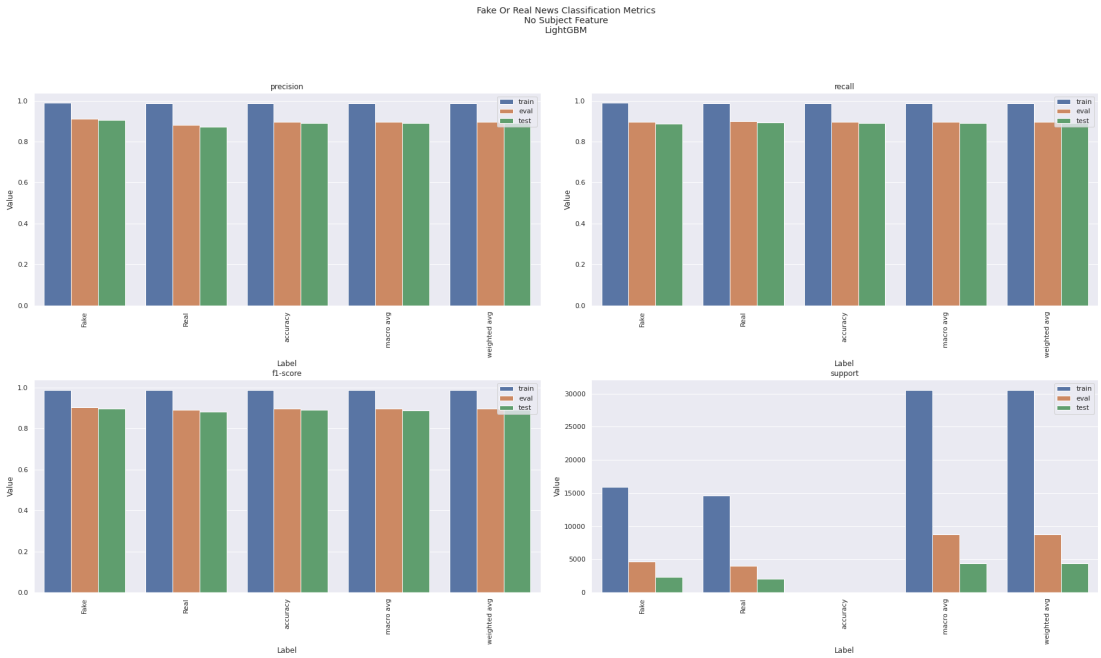

Figure 5: Fake Or Real News Classification Metrics. Using LightGBM Classifier on pre-processed extracted features.

4 Conclusion

We implemented and LLM as a way to extract intermediate features for a downstream task that are more interpretable than typical black box DNN algorithms. We find that the hypothesis of emotional sentiment analysis and the authenticity of news to be strongly correlated appears to hold true (at least for the tested dataset) and would be practical from a performance standpoint to deploy. Additionally we provide an example of how this intepretability may look in the *Analyzing Articles With Intepretability* section in the appendix. Future steps include cutting down on LLM model size and inference latency and deploying this framework as a browser extension.

With a method that intervenes in the spread of disinformation while still being interpretable by the user we hope to help stop this spread while building a sense of trust with the users themselves.

5 References

[1] Peng, H., Romero, D. M., & Horvát, E.-Á. (2022). Dynamics of cross-platform attention to retracted papers. *Proceedings of the National Academy of Sciences*, *119*(25). https://doi.org/10.1073/pnas.2119086119

[2] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.372

[3] Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

[4] Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).

6 Appendix

Fake or Real News Classification Metrics

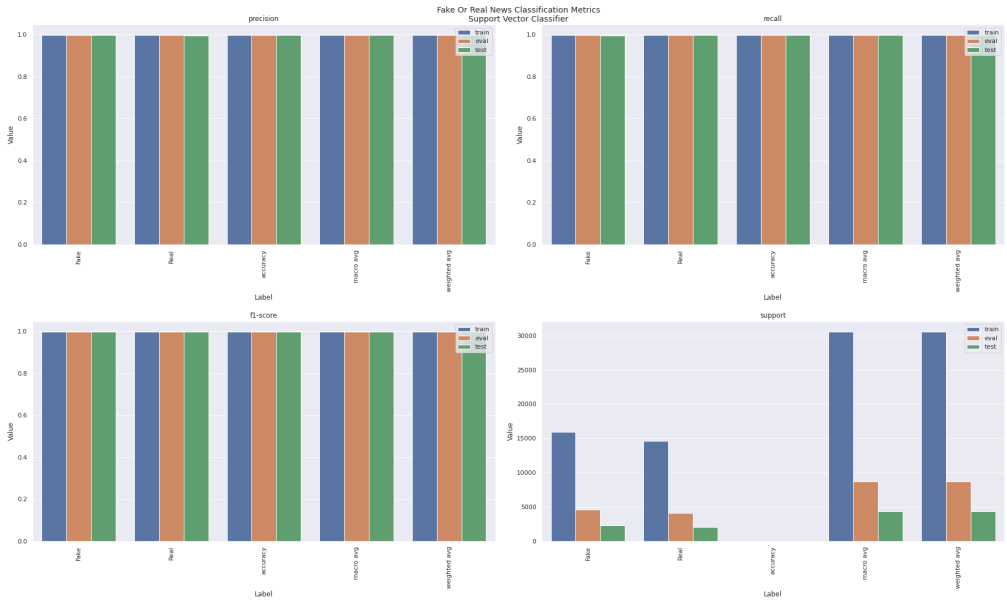

Figure 6: Fake Or Real News Classification Metrics. Using Support Vector Classifier on original extracted features.

GoEmotions Confusion Matrix

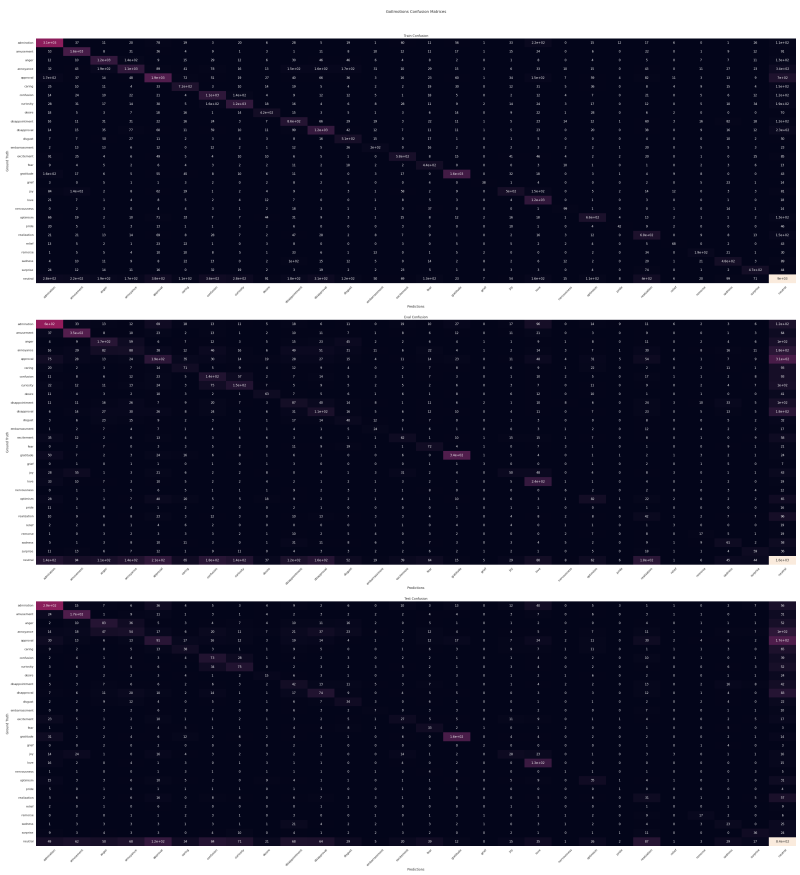

Figure 7: Confusion Matrices on train, evaluation and test samples from GoEmotions dataset using DistillBERT.

T-SNE Embeddings For Fake Or Real News Extracted Features

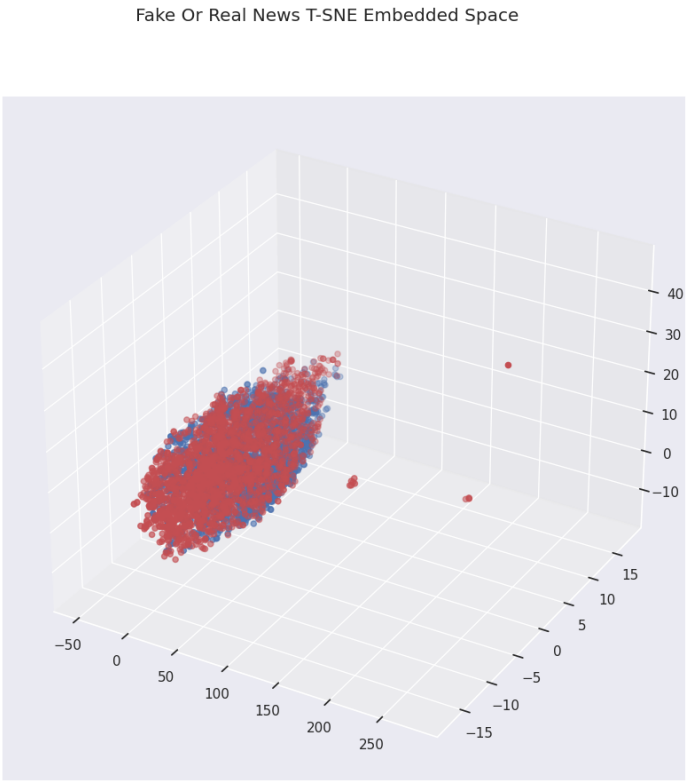

Figure 9: Fake Or Real News TSNE Embeddings on original extracted features.

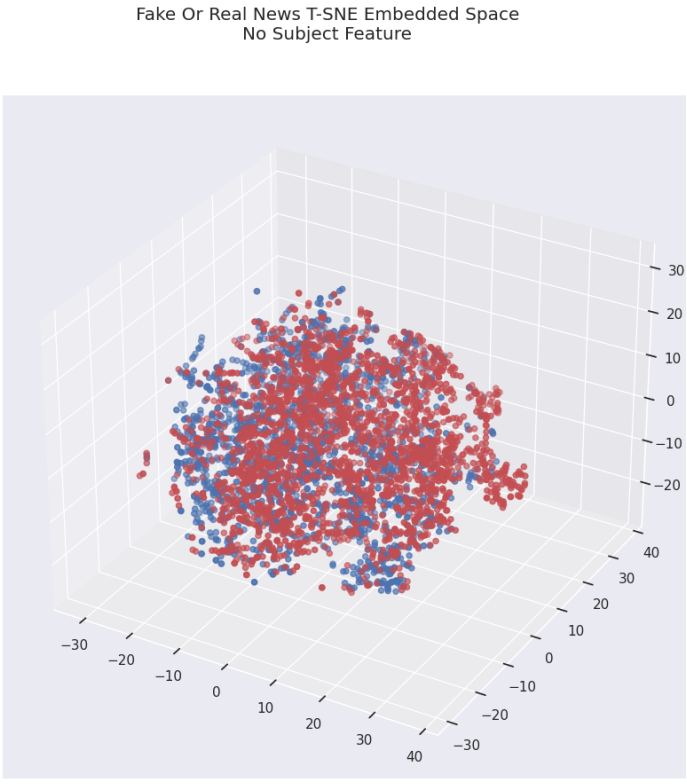

Figure 8: Fake Or Real News TSNE Embeddings on preprocessed extracted features.

Correlation For Fake Or Real News Extracted Features

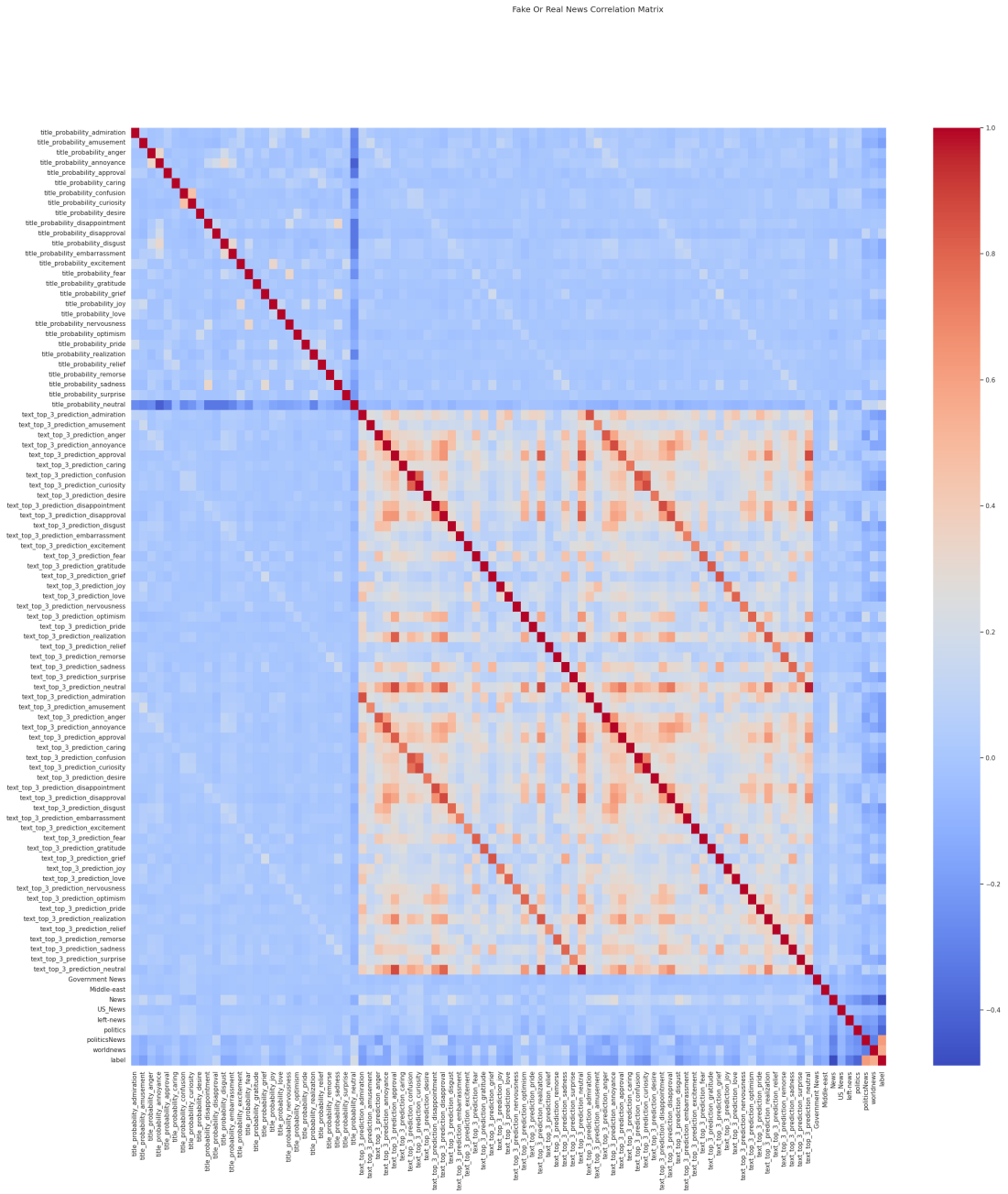

Figure 10: Fake Or Real News Correlation Matrix on original extracted features.

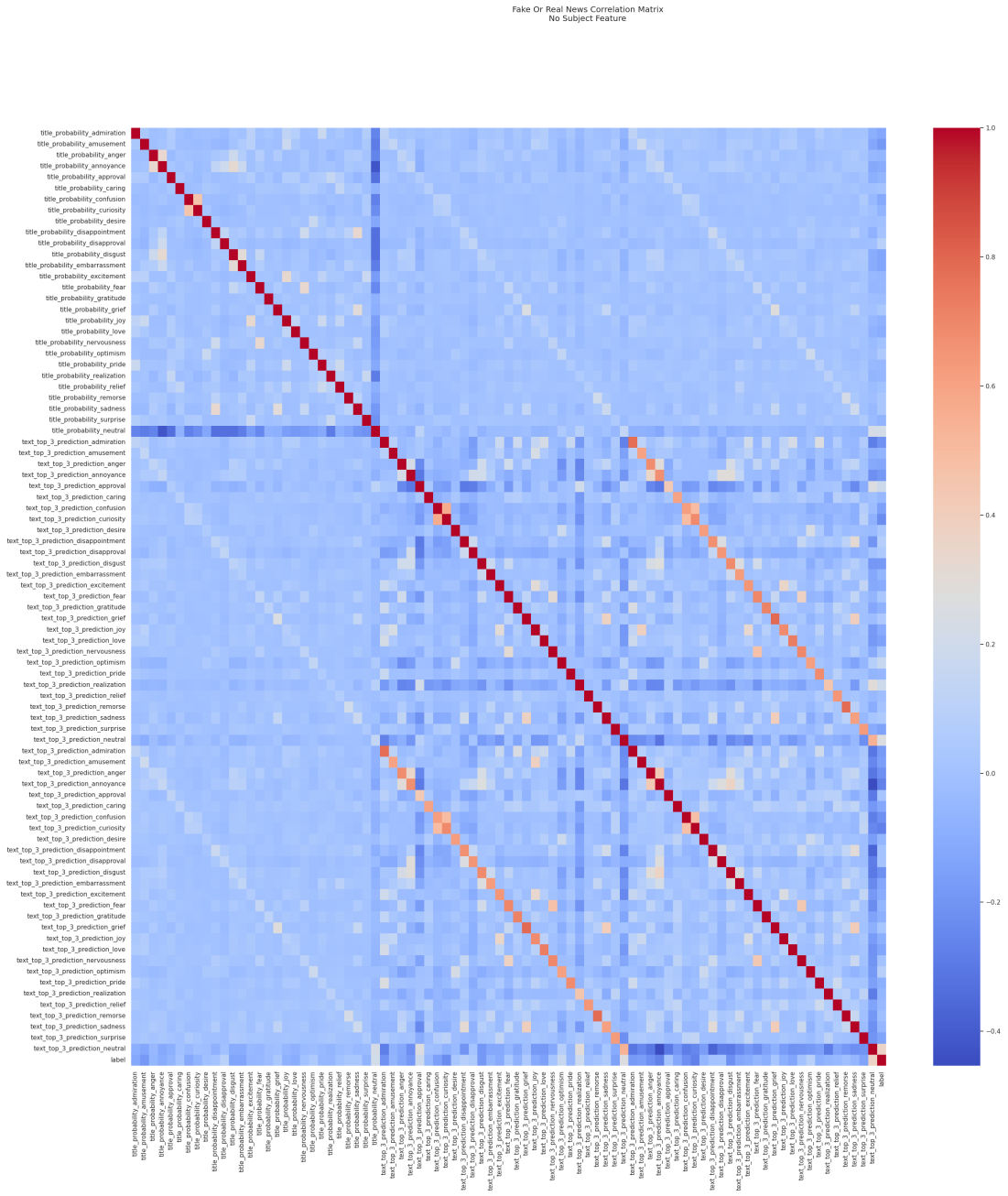

Figure 11: Fake Or Real News Correlation Matrix on preprocessed extracted features.

Confusion Matrices For Fake Or Real News Dataset

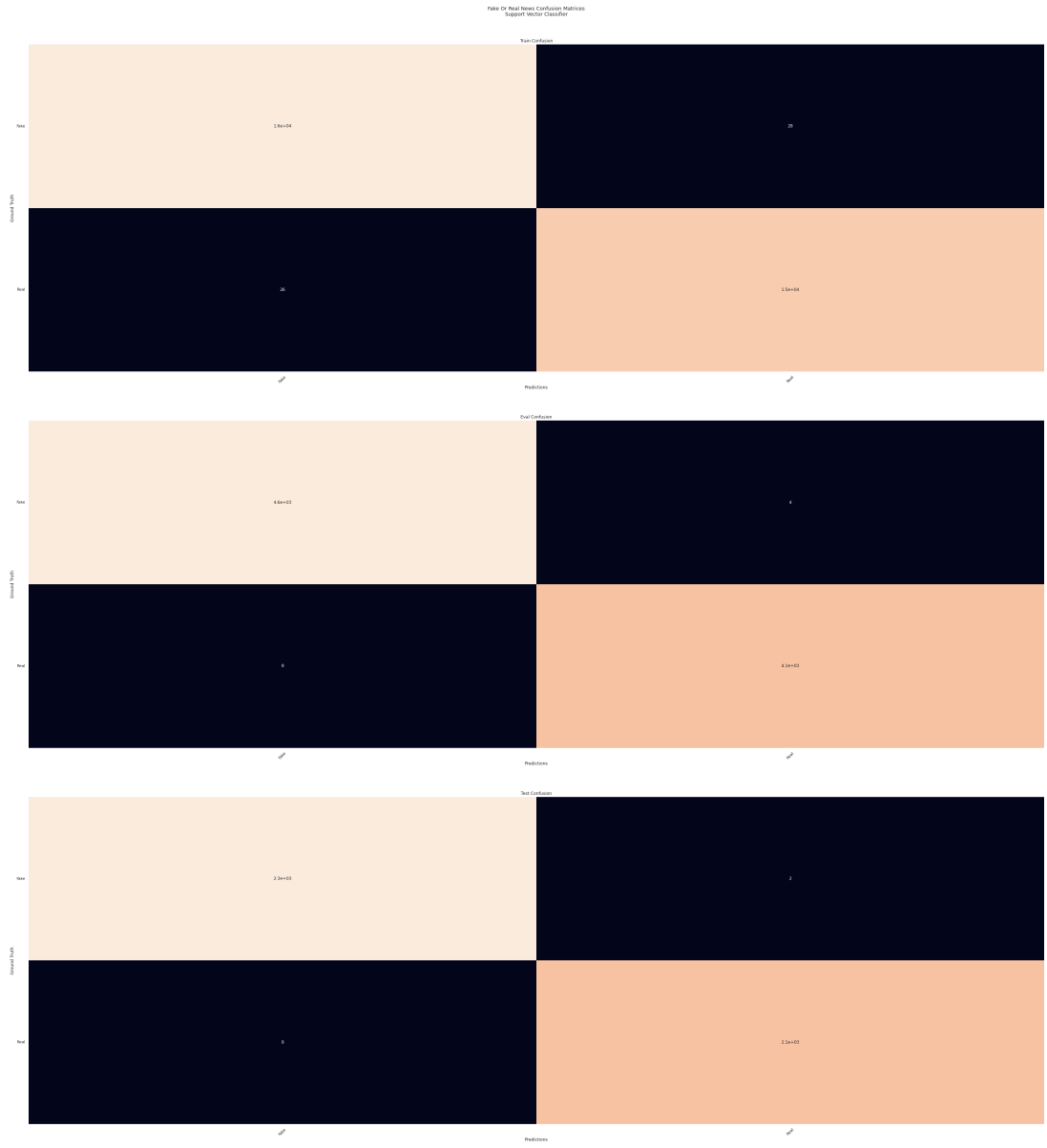

Figure 12: Fake Or Real News Confusion Matricies from Support Vector Classifier on original extracted features.

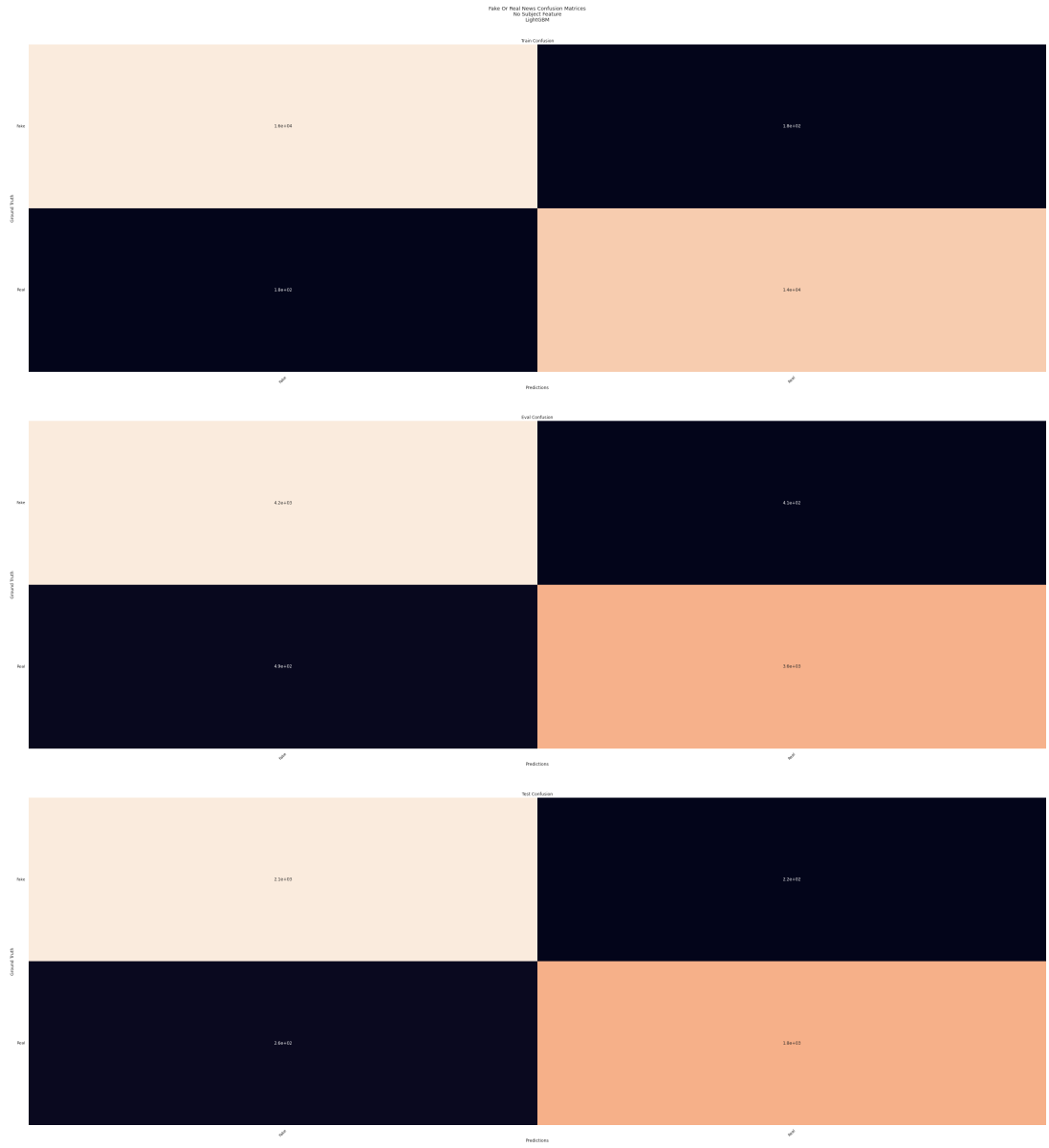

Figure 13: Fake Or Real News Confusion Matricies from LightGBM Classifier on preprocessed extracted features.

Analyzing Articles With Intepretability

Article 1

<neutral: 86, disapproval: 6, realization: 3> **HUD OFFICIAL Spends $366,000 in Fed Funds on Booze, Makeup, Clothes and Homes…You Won't Believe Her Punishment!** </>

<confusion: 46, neutral: 33, curiosity: 15> WHY IS THIS CRIMINAL ALLOWED TO JUST SKATE BY WITH SUCH A SHORT AMOUNT OF JAIL TIME?We know that this woman is working on a plea bargain but FIVE YEARS??? </>

<approval: 44, annoyance: 13, anger: 11> The two things that are so outrageous about this entire case are that there was NO OVERSIGHT and that this WENT ON FOR 8 YEARS!! </>

<neutral: 37, disapproval: 14, anger: 9> !USING THE FEDS AS YOUR OWN PERSONAL PIGGY BANK:Lorena Loren s job was to help poor people find housing Instead, prosecutors say, she was stealing from the poor to buy booze, makeup, clothes and homes for her relatives.But after years of scheming, authorities say, the government caught on. </>

<neutral: 89, realization: 8, approval: 1> IT TOOK 8 YEARS!In a public corruption case that surfaced in federal court today, Loren, the former executive director of the St. Clair Housing Commission, was charged with embezzling more than $300,000 in low-income housing funds and spending the money on herself and family.According to a charging document filed in U.S. District Court, Loren s scheme ran between 2008-2016 and benefited many family members, including her husband, son and son-in-law.Prosecutors say Loren s criminal activity started in 2008 five years after she was appointed executive director of the St. Clair Housing Commission, which administers federally funded low-income housing programs in St. Clair County.According to court records, Loren used her job in many nefarious ways, including: She fraudulently used the housing commission s two credit cards to buy personal items for herself and relatives, including nearly $135,000 in online purchases from Amazon.com; $14,364 in purchases from Sam s Club in Port Huron, and $16,460 in purchases from various Walmart stores. </>

<neutral: 95, approval: 1, optimism: 0> The purchases included, among other things, adult and infant clothing, furniture, food, beauty supplies, medications and alcoholic beverages. </>
<disapproval: 48, realization: 14, approval: 12> Her unauthorized Amazon.com purchases included $60,000 worth of bedroom furniture, mattresses, appliances and other household items for family members. </>

<neutral: 97, approval: 1, realization: 0> The items were shipped to Loren s relatives in Georgia and Florida. </>

<neutral: 84, realization: 7, confusion: 2> She embezzled $24,600 in commission funds to pay for her son s rental unit. </>

<neutral: 84, realization: 11, disapproval: 1> She lied on four government housing contracts by claiming that her son-in-law was the landlord of a rental property for low-income residents in Michigan, when it was really her home in Port Austin. </>

<realization: 38, neutral: 15, disapproval: 12> She falsified lease agreements and stole public housing money that was supposed to help poor people to rent a home for her son. </>

<neutral: 95, caring: 2, approval: 0> She directed members of her family to establish bank accounts so that federal rental subsidy payments could be deposited into those accounts. </>

<neutral: 99, realization: 0, approval: 0> The relatives then used the money for their own personal use, including rent for their Florida residences.Loren is charged in what is known as an information, which typically means she is working on a deal with the government. </>

<neutral: 79, anger: 7, disapproval: 5> She is charged with conspiracy to commit federal program fraud. </>

<neutral: 85, realization: 7, optimism: 1> If convicted, she faces up to five years in prison.Loren home Lorena Loren bought this $325,000 home in Georgia after federal prosecutors say she embezzled more than $336,000 from the St. Clair Housing Commission.JUDICIAL WATCH: HUD s inspector general has testified before Congress about the severe mismanagement at offshoots functioning independently yet federally funded in states across the country that have fleeced taxpayers out of hundreds of millions of dollars. </>

<realization: 58, neutral: 38, surprise: 0> He testified that his office discovered more than $200 million in questionable spending at local public housing agencies (PHAs) since 2012. </>

<neutral: 95, disapproval: 1, realization: 0> PHAs are created by states, operate at the city or county level and administer the federal program known as Section 8 to provide low-income people with affordable housing. </>

<neutral: 72, realization: 14, confusion: 2> Many of these local public housing agencies are run by people with troubled backgrounds that somehow manage to remain in high-ranking positions at the agencies, the watchdog told lawmakers back in 2014.DR BEN CARSON NEEDS TO CLEAN THIS UP STAT!READ MORE: DFP </>

Prediction: Fake: 95 Label: Fake

Article 2

<neutral: 99, approval: 0, sadness: 0> **Saudi-led coalition to reopen Yemen's Hodeidah port, Sanaa airport for aid** </>

<neutral: 67, fear: 22, anger: 1> DUBAI (Reuters) - The Saudi-led military coalition fighting Houthi rebels in Yemen said on Wednesday it would allow humanitarian aid access through Yemen s port of Hodeidah and United Nations flights to the capital Sanaa, more than two weeks after blockading the country. </>

<neutral: 68, disapproval: 21, annoyance: 2> The coalition closed air, land and sea access to the Arabian Peninsula country on Nov. 6 to stop the flow of arms to the Houthis from Iran. </>

<neutral: 98, realization: 0, approval: 0> The action came after Saudi Arabia intercepted a missile fired toward its capital Riyadh. </>

<disapproval: 44, neutral: 39, annoyance: 4> Iran has denied supplying the Houthis with weapons. </>

<neutral: 48, disappointment: 19, disapproval: 7> Soon after the closure, U.N. aid chief Mark Lowcock warned that the blockade could spark the largest famine the world has seen for many decades with millions of victims unless the coalition gave access to humanitarian aid. </>

<neutral: 98, approval: 0, realization: 0> The Saudi-led coalition said in a statement on Wednesday that from Nov. 23 the Red Sea port of Hodeidah would be reopened to receive food aid and humanitarian relief, and Sanaa airport would be open for UN flights with humanitarian relief. </>

<neutral: 75, approval: 9, annoyance: 1> We re monitoring these developments, U.N. spokesman Farhan Haq told reporters in New York. </>

<approval: 72, neutral: 14, excitement: 7> If that were to happen that would be a very welcome and critically important development. </>

<neutral: 94, realization: 2, approval: 1> We made clear the tremendous amount of needs on the ground, Haq said. </>

<neutral: 95, approval: 2, optimism: 0> Earlier this month the coalition said it would allow aid deliveries through the government-held port of Aden. </>

<neutral: 99, realization: 0, confusion: 0> However, around 80 percent of Yemen s food imports arrive through Hodeidah. </>

<neutral: 80, sadness: 13, fear: 1> The United Nations has said some seven million people in Yemen are on the brink of famine and nearly 900,000 have been infected with cholera, a waterborne disease that causes acute diarrhea and dehydration. </>

<neutral: 85, approval: 3, optimism: 3> Aid groups said there also needed to be commercial access to Yemen for food and fuel shipments. </>

<neutral: 87, realization: 6, disapproval: 4> Humanitarian aid alone cannot meet the needs of Yemenis who are unjustly bearing the brunt of this war, Paolo Cernuschi, Yemen country director at the International Rescue Committee, said in a statement. </>

<neutral: 94, approval: 2, desire: 1> Jan Egeland, secretary general of the Norwegian Refugee Council and a former U.N. aid chief, posted on Twitter, We need all ports to open and access for commercial food and supplies to large civilian population. </>

<neutral: 86, disapproval: 9, realization: 3> Humanitarian aid alone cannot avert hunger. </>

<neutral: 86, caring: 9, approval: 0> The International Committee of the Red Cross (ICRC) said on Wednesday that it had evacuated from Sanaa five staff members in need of urgent medical assistance. </>

<neutral: 71, anger: 6, fear: 6> The Saudi-led coalition has been targeting the Houthis since they seized parts of Yemen in 2015, including the capital Sanaa, forcing President Abd-Rabbu Mansour Hadi to flee. </>

<neutral: 92, disgust: 2, fear: 0> The Houthis, drawn mainly from Yemen s Zaidi Shi ite minority and allied with long-serving former president Ali Abdullah Saleh, control much of the country. </>

Prediction: Real: 99 Label: Real