MORALITY ANALYSIS ON TWEETS FOLLOWING TERRORIST ATTACKS USING NATURAL LANGUAGE PROCESSING

Muyan Li
Moscrop Secondary School
MehtA+Tutoring

Siddarth Kappa
South Brunswick High School
MehtA+Tutoring

Danny Liu
Trinity College School
MehtA+Tutoring

Hana Hashmi
Smithtown High School East
MehtA+Tutoring

July 29, 2021

ABSTRACT

Previous research has shown that after a terrorist attack, people demonstrate an increase in salience of moral institutions such as the respect for authority and purity due to the need for social cohesion. Using sentiment analysis and machine learning classification models on Twitter tweets after the El Paso terrorist attack and Las Vegas terrorist attack, we intend to extend the previous research on a bigger database in this research study. After hand annotating 900 tweets with each of five morality, we use Support Vector Machine (SVM) to classify the tweets into 5 moral intuitions. Comparing the percentage of tweets that express each moral foundation category, we find that our comprehension of the El Paso attack is aligned with or similar to the previous studies, but our conclusion of the Vegas attack does not. This paper documents how we utilize machine learning models to classify people's morality and record our findings on the sentiment of tweets after a terrorist attack.

1 Introduction

On October 1, 2017, there was a terrorist attack at a music festival in Las Vegas, causing 61 deaths altogether. On August 3, 2019, there was another terrorist attack within a Walmart located in El Paso, Texas. 23 people were killed and 23 more were injured. Experiencing the vast impact of terrorist attacks and the quick spread of information through the media, we are curious about how those attacks can influence people and want to utilize machine learning algorithms to measure how twitter users' moralities change after terrorist attacks. Since there isn't a universal standard to measure a reaction, we decide to classify the tweets by using the five moral sensitivities stated in the moral foundations theory. These 5 sensitivities are care, fairness, in-group loyalty, respect for authority, and purity. (1) Care refers to the compassion in the tweet, and it applies when the emotion of empathy is shown. (2) Fairness calls upon the idea of justice, rights and autonomy. (3) In-group loyalty represents the act of self-sacrifice and the love to the community a person belongs to. (4) Respect for authority demonstrates the virtue of leadership and followership and the nature to obey social or societal traditions. (5) Purity signifies our notion to strive for more noble and elevated lifestyles. After classifying the tweets from the 2019 El Paso attack and the 2017 Vegas attack, we split our data into immediate responses, which is within 18 hours the attack had happened, and later responses, which is after the attack had happened 18 hours. Comparing these two groups, we analyze whether people's morality increases after the attack has occurred.

2 Related Work

When we started this project, our task was to extend the research of this paper [1] which was about testing and experimenting with the five moral sensitivities. The authors of this paper conducted an experiment that took participants and showed them images and videos of terrorist attacks. The participants were then presented the opportunity to

donate to charity. The results showed that after being exposed to media about terrorist attacks, participants were more likely to donate to charity, concluding that participants were more empathetic. The authors of this paper measured the participants' reactions by observing how a participant's moral sensitivities changed, which gave us the inspiration to measure reactions using the five moral sensitivities.

While researching this topic, we found a paper [2] that was tackling a very similar problem as we were. The authors wanted to use machine learning models to classify tweets under moral sensitivities so they could have an insight into phenomena such as protest dynamics and social distancing. They gathered tweets from relevant topics such as Black Lives Matter, the 2016 Presidential election, and Hurricane Sandy. The authors also acquired around 35,000 annotated tweets, which are labelled by professional annotators, and had the hope of other researchers using their tweets and continuing this topic so other methods and machine learning models could be developed. Unfortunately, when we extracted the tweets and tried to utilize it as our training data, around 95% of the tweets were either deleted or not accessible anymore, so there was only a small amount of tweets left that could be used. However, we still did not end up using these tweets because the data are extremely unbalanced: there are many tweets that are classified as in-group loyalty but few tweets that are classified as care or purity.

3 Methodology

3.1 Dataset

The dataset that we used was a Twitter corpus and many tweets were scraped from the time around the shootings happened. The 2019 El Paso attack contains 41, 071 tweets, and after we split the data into immediate responses and later responses, 18, 415 or 45% of the tweets belong to the immediate responses, and 22, 656 or 55% of the tweets belong to the later responses. The 2017 Las Vegas contains 6, 024 tweets, and after we split the data, 3, 516 or 58% of the tweets belong to the immediate responses, and 2, 508 tweets or 42% of the tweets belong to the later responses. We choose 18 hours as our splitting point because it is when our data splits into approximately half. Since we do not want any of the groups to be under-represented or over-represented, we choose to split our data in approximately half. In addition, we do not have access to data before the attack, so we cannot compare the percentage of tweets that contain each moral value before and after the attacks and make conclusions about whether each virtue increases in percentage. Therefore, we choose to split our data at the 18 hours benchmark after the attack happened to investigate whether the percentage of tweets that express each virtue continues to increase. This is an extension to previous research questions on real-time data.

Preprocessing: the tweets within the Twitter corpus had come in a .json file, which we could not access and convert it into a dataframe easily. Therefore we first converted the file into a .csv file and then dropped unnecessary data such as user ID. There were some tweets that were in different languages, so we got rid of all tweets that weren't English. Now we had all of the data we needed, and it was easily accessible.

3.1.1 Data Augmentation

For the tweets that were extracted, we manually labelled the categories they fall under, and each tweet could potentially fall under more than one category. El Paso had 500 annotations, and Las Vegas had 400 annotations. Due to our limited time, a total of 900 tweets is what we can manually label. In addition, because we are neither professional annotators nor humanities researchers, our labelling only reflects our own opinions, which could be biased. Therefore through someone's perception, the model can be perfectly correct, but through another person's perception, the model can be completely wrong. These 900 tweets we labelled act as our training and testing data. For all our models, we did an 80/20 split, which means that about 720 tweets will act as training data, and 180 tweets will act as testing data.

3.1.2 Extracting Features and Labels

The features we used are similarities between words based on Word2Vec algorithm, the length of the tweet, the number of hashtags, and TF-IDF. To extract these features, we utilize the moral foundation dictionary, which is a collection of words researchers classified as having high association with each category. For example, the care category will contain words such as "love" and "caring". Furthermore, we also tried to utilize the moral foundation dictionary 2.0, which is an upgraded version of the first dictionary, containing significantly more words for each category. Below is a detailed description of how we extract each feature.

Word2Vec: Word2Vec is a method to vectorize a word in a text, create an embedding for a word, and return the similarity between different words. To extract our features, we utilize a pre-trained Word2Vec model on google

news and run it on each moral sensitivity's dictionary. We then return the number of words in a tweet that have more than 0.25 similarity with each word in the moral foundation dictionary and the maximum similarity between a word in a tweet and a word in each categories' dictionary.

Length of Tweet: the length of a tweet is simply the total number of words that were in the tweet.

Number of Hashtags: Hashtags are one of the special features social media has. Therefore, we include the number of hashtags as one of our features.

TF-IDF: TF-IDF, or Term Frequency-Inverse Document Frequency, vectorizes a text and determines how important a word is to a corpus. It weighs the collection of words in the corpus, and doesn't give huge weightage to stop-words such as "and", "the".

For each tweet in our training and testing data, we utilize one-hot vector to convert our label to a matrix, with each line of the matrix representing whether the tweet is classified as expressing the intuitions. For example, if a tweet is classified as 1 for care, it means that this tweet is expressing care. However, since the features for each tweet are distinct for each category, we have to run the model 5 times, iterating through the 5 categories.

3.2 Model

The models we used were support vector machine (SVM), SVM 2.0, and k-nearest neighbors (KNN).

Model 1:
In the first approach, we used SVM as our classification model. We made use of the first moral foundations dictionary to calculate the similarities between the tweet and distinct virtue or intuition. Iterating through the 5 moral intuitions one at a time, we concatenated the results and generated a 5 by 1 matrix as a label for each tweet. After experimenting around with the various parameters, we set the regularization parameter C to 10 and the kernel parameter to linear.

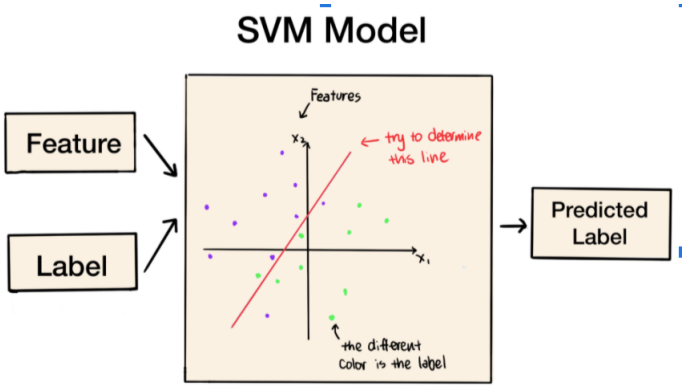

Figure 1: Visualization of SVM

Model 2:
SVM 2.0 was a replica of the first model, except we used the moral foundation dictionary 2.0 instead of 1.0. However, due to the massive set of words we ran into problems because the model took 30 hours to extract the features. We do not have enough time for the model to continuously run for 30 hours, and our RAM is breaking due to our limited resources. Therefore, we were not able to finish this model and obtain the accuracy.

Model 3:
We used the first moral foundations dictionary to acquire features for this KNN model, which is similar to model 1. Furthermore, we tested K, the number of nearest neighbours, from 1 to 30 and determined which K has the best accuracy for each category. For care, k = 28; fairness, k = 7; loyalty, k = 10; authority, k = 7; and purity, k = 16.

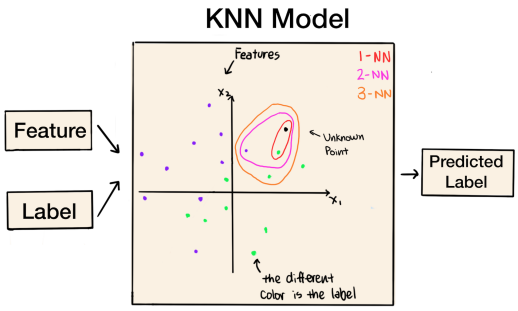

Figure 2: Visualization of KNN

3.3 Accuracy

Table 1, Figure 3, and Figure 4 records the accuracy of model 1 and model 3. We can see that the KNN model has a slightly better accuracy than the SVM model, but overall, they have approximately similar results. Due to the fact that the SVM model is much faster to run, we chose the SVM model to label our dataset, as it is more convenient for us.

	Average	Care	Fairness	Loyalty	Authority	Purity
SVM	75.1%	67.8%	72.2%	83.3%	81.7%	70.6%
KNN	79.89%	68.33%	77.78%	90.00%	86.11%	77.22%

Table 1: Model Accuracies

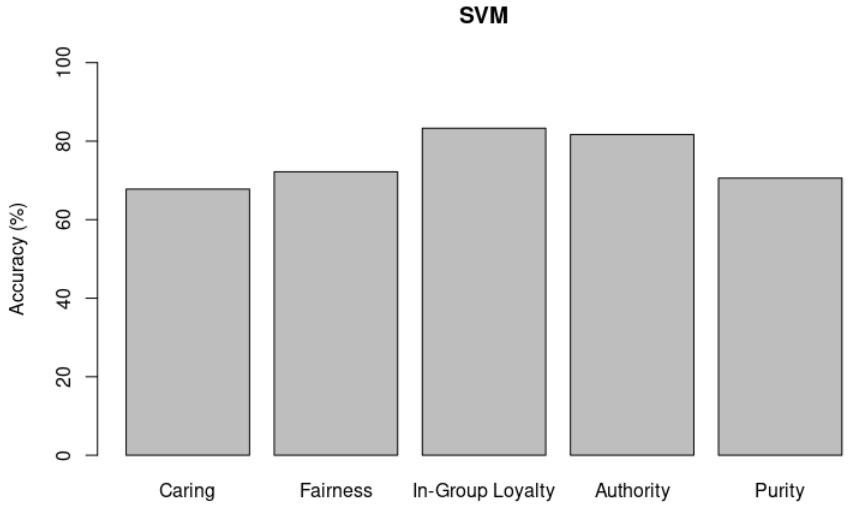

Figure 3: SVM Accuracy

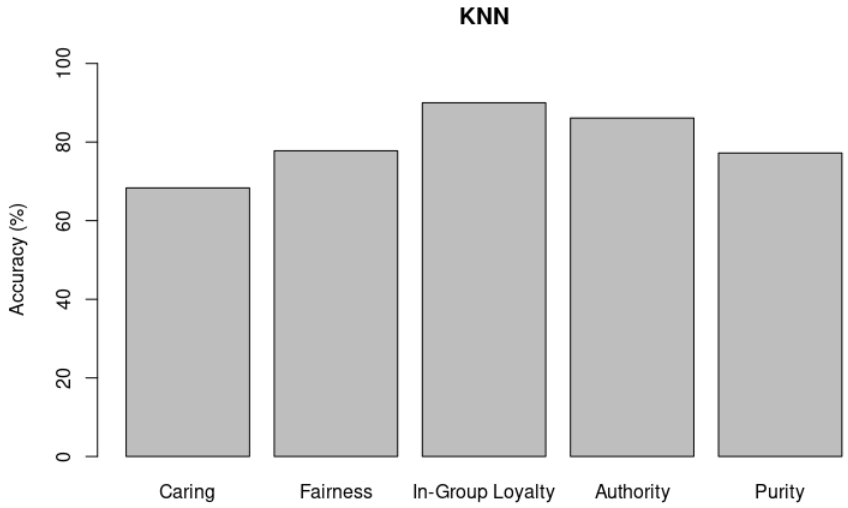

Figure 4: KNN Accuracy

4 Results

After splitting our two datasets of El Paso and Vegas into immediate and later responses, we calculate the percentage of tweets that express each virtue in these two groups. Our analysis results are showcased in the tables below.

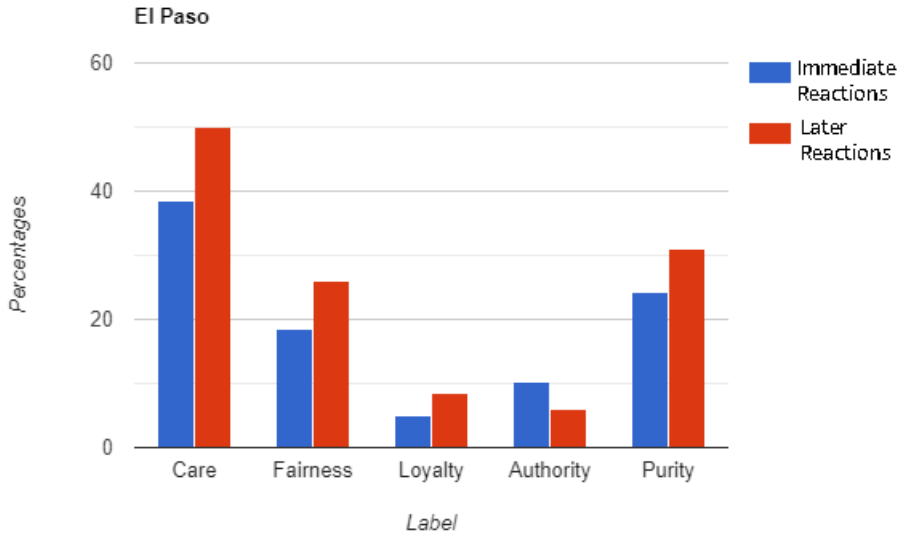

Figure 5: El Paso Graph Results

When we compare the El Paso and the Vegas dataset as a whole, we can see that care is the virtue people express the most after a terrorist attack, followed by purity and fairness. This shows that many tweets show support to the victims and try to comfort the people that are harmed. However, when we compare the immediate and later responses within each attack, El Paso and Las Vegas attacks give very different results. For the El Paso attack, every intuition except respect for authority demonstrates an increase in percentage, which is consistent with previous research findings.

ElPaso:

	Care	Fairness	Loyalty	Authority	Purity
Before 18h	38.4%	18.5%	4.89%	10.3%	24.3%
After 18h	50.0%	26.0%	8.42%	6.08%	31.0%

Table 2: El Paso Table Results

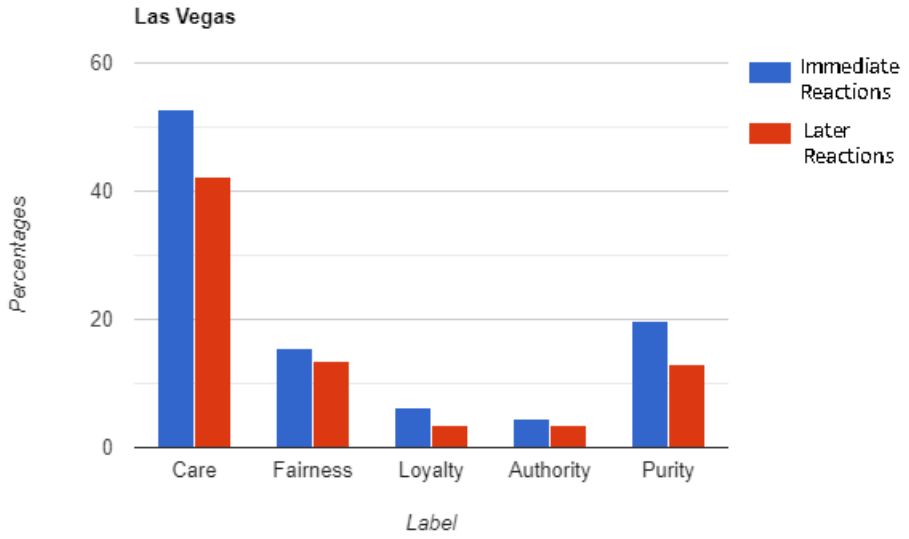

Figure 6: Las Vegas Graph Results

Vegas:

	Care	Fairness	Loyalty	Authority	Purity
Before 18h	52.8%	15.4%	6.22%	4.44%	19.8%
After 18h	42.2%	13.5%	3.50%	3.50%	12.9%

Table 3: Las Vegas Table Results

However, for Las Vegas, every moral intuition demonstrates a drop in percentage. This inconsistent result may be the result of the limited dataset and the different nature of the Las Vegas attack. To begin with, compared to the 41,000 tweets we have for the El Paso attack, we only have about 6,000 for the Las Vegas attack. This may cause our predicted labels to be biased and skewed. In addition, since the attack happened in 2017, a lot of the real-time tweets have been deleted, resulting in our limited dataset. Moreover, the El Paso attack is both domestic terrorism and a hate crime, but for the Vegas attack, it is only domestic terrorism. These two different portrayals by the media and classifications or nature of the attacks may also contribute to different reactions people have or tweets people posted, therefore resulting in the inconsistent results we obtained.

5 Conclusion and Future Work

Our classification of tweets using natural language processing models are generally consistent with previous research's results. In the future, we hope that there could be an improvement in the efficiency of the different models, elevating the accuracy. In addition, future works can be done with tweets labeled by trained annotators. Moreover, future researchers could investigate how the location of attacks and how the different nature of attacks or the different portrayals of the attack on media influence people's morality.

6 Division of Labor

We divided the work as follows:

- Muyan (Anna) Li - Lead Programmer, Annotator, Writer
- Siddarth Kappa - Programmer, Annotator, Writer
- Danny Liu - Programmer, Annotator, Writer
- Hana Hashmi - Programmer, Annotator, Writer

7 Acknowledgements

We would like to thank Ms. Andrea Jaba, Ms. Haripriya, Mr. Bhagirath, and Mr. Mohammad Sharafat for their assistance and support on this project. Special thanks to Professor Lindsey Hahn for providing us with the dataset and giving us support.

References

[1] Ron Tamborini, Matthias Hofer, Sujay Prabhu, Clare Grall, Eric Robert Novotny, Lindsay Hahn, and Brian Klebig. The impact of terrorist attack news on moral intuitions and outgroup prejudice. *Mass Communication and Society*, 20(6):800–824, 2017.

[2] Joseph Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida M Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, and et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment, Apr 2019.