# **CS791 Final Report**

# **The EPIK Project**

**Yunruo Ni**

**Quincy Heon**

Table of Content

Introduction-----------------------------------------------------------------------------------3

Methods-------------------------------------------------------------------------------------3-5

Results---------------------------------------------------------------------------------------5-6

Discussion----------------------------------------------------------------------------------6-8

Conclusion------------------------------------------------------------------------------------8

Appendix----------------------------------------------------------------------------------8-11

# **Introduction**

The EPIK project is dedicated to disrupting the online sex trafficking industry by first drawing in potential offenders with fake ads on the dark web and then speaking to responders, either through trained volunteers or chat bots. A successful conversation is one in which the conversation reaches the point of offering resources (bronze standard), the buyer accepts these resources (silver standard) or is convinced enough by the conversation to not reoffend in the future (gold standard). The client has supplied us with records of these interactions and wants us to determine if there are strategies volunteers can take to have a better chance of success, and if there are mistakes being made by volunteers that are more likely to lead to failure.

# **Methods**

This project involved three main methods: data engineering, sentiment analysis and classifications model training.

At the beginning of our project, a significant amount of the work done has been data filtering and condensing. There are multiple CSV files with thousands of rows and dozens of columns, some of which are over different time periods or only record certain aspects of the conversation. Most files have a number of either redundant or unhelpful columns of data, and some of the important columns have missing entries, most notably the “accepted resources” column. So, before going much further, it was necessary to reduce this data set into something workable. None of the phone conversations had transcripts, and we wanted to focus on things we can change, so we decided to focus our energy on investigating text conversations between buyers and trained volunteers.

Although it isn’t recorded anywhere, it is easy to find which conversations were a buyer’s final offence using the unique buyer IDs and the date and time of the conversation. Naturally, the latest timestamp conversation for each unique buyer ID is the conversation that convinced them not to call back (at least in theory). For our other “success” condition, the client’s system records when a buyer accepts resources, but doesn’t always record a failure, leaving the spot blank instead. These blanks we will consider failures for the purpose of this study. Fortunately, the system was good about recording when a buyer was offered resources.

Once a suitable dataset was acquired, we moved on into the sentiment analysis of the text conversations. Because the datasets of this project are unsupervised data, we can not use the natural language process prediction model to predict the sentiment of messages. We chose to use existing sentiment analysis libraries to evaluate our text messages. The first industry sentiment analysis we applied is “NLTK VADER Sentiment Analysis”. We choose the compound score of the VADER library where positive scores show positive sentiment in sentences. To decrease the potential bias by NLTK library, we also employed Textblob Sentiment Analysis into our text messages. The sentiment function of textblob returns two properties, polarity, and subjectivity. We added both of them into our analysis. In addition, we found that separating the agent and buyer’s contributions lead to more meaningful results, and so we found the above scores for each member in each conversation.

We then experimented with a number of models, including Random Forest, XGBoosting, and Decision Trees. We choose those popular classification models to test on our sentiment analysis along with other features. We roughly have 5744 paired data and split them into eighty percent of training datasets and twenty percent test datasets. According to the accuracy, the XGBoosting model has the highest accuracy with 0.5073. Except for seeking the high accuracy of prediction, we also want to find important features in the datasets. We used the exisiting function “feature importance” for all of these three models along with the additional permutation importance function for Random Forest Model. Lastly, we concluded the most common important features from those training results: “Purchase Attempt Numbers”, “Agent\_ConLen”, “Agent\_VaderSc”, “Agent\_Textblob\_Subjectivity”, “Visitor\_Textblob\_Polarity”.

Now with the sentiment scores and feature importance figures, we can begin to see the impact of the tone in the text conversations on success. We then used a Chi-squared significance test to determine if the success rate was significantly different when the agent was being negative versus positive over our gold, silver, and bronze success metrics.

# **Results**

We began with three success metrics, and for each we hypothesized that the agent’s sentiment score would significantly impact the odds of success. The corresponding null hypotheses, then, are that the scores would not significantly impact the odds of success. We chose α=0.05 for our significance level, and we only have two rows (success or failure) and two columns (negative or positive agent sentiment) so we have one degree of freedom, giving us a Chi-squared test statistic of 3.84. The Chi-squared values from our actual data gave 1.62, 5.60, and 29.98 for our gold, silver, and bronze standards, respectively. We can reject the null hypothesis if our chi-squared value from the data is larger than the test statistic, and so we can reject the null hypothesis for both our silver (buyer accepting resources) and bronze (offered resources) standards. The data does not allow us to reject the null hypothesis for the gold standard (buyer never re-offends) at our chosen significance level.

We now know that it is unlikely (less than one in twenty) that the differences in success rates between the positive and negative sentiment interactions is not significant. Without the double negatives, we can say that we’re at least 95% sure that sentiment has a real impact on success. The chart in the appendix shows the percentage success rates for our three success metrics, divided into positive and negative sentiment. We can see that positive sentiment always has a higher success rate within a metric than the negative sentiment. It is worth noting that although the gold standard shows the same trend as the silver and bronze, the difference is not statistically significant, at least within the confines of this study. For silver and bronze, however, we can confidently say that positive agent sentiment is more successful.

# **Discussion**

Our results show that agents who are being nice have a higher chance of success than agents who are not. This is true in our results for all three standards, but we can only statistically reject the null hypotheses for the silver and bronze standard, as the effect wasn’t pronounced enough in the gold standard to make a conclusion. This doesn’t necessarily mean that agent sentiment has no impact on buyer call-backs, but it does mean that more studies will be required to prove it. For the silver and bronze standards, we can conclude that more agent positivity leads to more success in accepting and offering resources.

Probably the largest challenge was getting the sentiment analysis working. These texts are very unusual in their subject matter, and there are words that would normally have a certain connotation in everyday speech, but in this context have a different connotation. A good example is profanity, which normally implies anger or frustration but in this case is sometimes used neutrally, as a descriptor instead of an outburst. In addition, the texts are totally unlabeled for their sentiment, and manually reading and tagging even a subset would take a long time, and we wanted to limit our exposure to the explicit texts as much as possible.

Unfortunately, our results were not enough to conclude anything about the gold standard. We also limited our study to the text conversations between real buyers and volunteers, and so our conclusions can only really be applied to these groups. A future study would be needed to review the gold standard and confirm the trend more concretely.

Besides the sentiment information impact on final successes, some other features also showed impact on results. The “Agent\_ConLen” feature showed a positive relationship with y-labels. When the case with higher agent conversation length, the possibility of their results being positive will increase. (For example, when we just choose the top 100 agent conversation, those conversations have more labels 3,4,.) However, the feature “Total\_Visitor\_Wait” (visitors’s waiting time during the conversation)shows a negative relation with the result. While doing simple regression on “Total\_Visitor\_Wait” and “y\_2classes”, the coefficient of “Total\_Vistior\_Wait” is negative.

**Conclusion**

In this project, we used machine learning techniques along with statistical methods to give an evaluation on the EPIK Program.

According to both machine learning models’ feature importances and statistical analysis, we can conclude that there are several main impact factors in the EPIK Program: Agent Sentiments, . Based on these factors, we give the following suggestions to the EPIK Program so that they may get better business outputs: 1)increase their agent sentiments; 2)let the agent try to make their sentence longer 3) decrease the buyers waiting time during conversation.

Besides suggestions on impact factors, we also have some advice on building their databases. Firstly, the data will be much clearer if the EPIK Program can set a explicit rule on defining success. Secondly, the EPIK Program adds more potential covariates into their datasets. While doing data engineering, we found that a large proportion of their data just talked about ID information. In addition, for sentiment analysis part, we rely on the existing library which gives label based social media messages. This makes our sentiment analysis bias since the EPIK Program has its own dialogue format. Results will be much more reliable if the EPIK Program can set a standard on labelling their data.

# Appendix1

|  |  |  |
| --- | --- | --- |
|  | Negative Agent Sentiment | Positive Agent Sentiment |
| Gold |  |  |
| Silver |  |  |
| Bronze |  |  |

# Appendix 2

|  |  |  |
| --- | --- | --- |
|  | Model Prediction Accuracy | Feature Importance from Model |
| Random Forest | 0.49956 | Coefficient Feature Importance  Permutation Importance |
| Decision Tree | 0.44212 |  |
| XGBoost | 0.5073 |  |