Project Betrayal Creative Extension

Alice Biryukov
alice.biryukov@epfl.ch

Quentin Juppet
quentin.juppet@epfl.ch

Claire Lefrancq
claire.lefrancq@epfl.ch

Abstract

The Linguistic Harbingers of Betrayal paper (Niculae et al., 2015) explores linguistic features that foretell the impending betrayal in a relationship between two people, using a dataset of communications between participants of the online strategy game Diplomacy. Following the work of the paper, we have implemented our own creative extensions that we discuss in this report. We have implemented a new feature, based on the support that the two players within a friendship provide each other in every season. Furthermore, by approaching the dataset from the standpoint of friendships rather than that of seasons, we were able to reach an accuracy of almost 60% on the predictions of whether a given friendship would end in betrayal.

1 Introduction

In our extension of the "Linguistic Harbingers of Betrayal" paper (Niculae et al., 2015), we reproduce the features used by the authors and use different machine learning models to compare their performance when predicting betrayal. We also consider new features: the amount of support provided by the victim and the betrayer, as well as aggregates of the features to characterize each friendship with their mean and variance. We ask ourselves whether these new features are relevant when trying to predict betrayal, and we train different machine learning models with them.

2 Related Work

The authors of the paper (Niculae et al., 2015) look at messages between pairs of people to explore whether betrayal has its warning signs in the language that the individuals use. The goal is to see if there are linguistic cues which are undetectable for the victim, but can be recognized by a machine. To this end, the authors use the communications between participants of an online war-themed strategy game (Diplomacy) that revolves around alliances and betrayals. They extract features from the written conversations such as sentiment, argumentation, politeness, or talkativeness. They compare those features with the occurrence of in-game acts of friendships or hostility to characterize the elements that might signal an impending betrayal. They then train a classifier based on their findings to predict betrayal and the time it will happen.

3 Dataset

We will be using the dataset provided with the "Linguistic Harbingers of Betrayal" paper (Niculae et al., 2015). It is a set of 500 friendships, half of them ending in betrayal. Each friendship contains a sequence of seasons. Within each season, the authors provide features extracted from the messages sent by each player, such as politeness, subjectivity, number of sentences, etc.

4 Methods

4.1 Reproduction of the Features from the Paper

To reproduce the features used in the paper, we computed the mean value of all the message features in a season, resulting in a list of season features. Some of the information about the messages (e.g. temporal features) was stored as a list of words, so we chose to replace them by the number of words in each list. When these lists were defined as N/A, we assumed that they contained no words, and thus assigned them a value of 0. In addition to the messages features, we also added two new features: the number of messages and the support: whether each player supported the other

during the season (1 if it they did and 0 if not). All of theses features exist twice, as they are a property of the player type: the victim or the betrayer.

4.2 Friendship Features

We considered features characterizing each friendship. To this end, we aggregated the features of all messages of each friendship, and computed their means and variances.

4.3 Training the Models

We selected three different models to train with our dataset and predict betrayal. We used 5-fold cross-validation as in the paper to select the model parameters, according to the accuracy.

4.3.1 Logistic Regression

Logistic regression is the model that was used by the authors of the paper for predicting whether a given friendship would end in betrayal or not. This model is based on a linear relationship. To implement the model, we used *Logit* from the *statsmodels* library.

4.3.2 Gradient Boosting Regressor

Gradient Boosting for regression is a nonlinear model that we have previously seen in a Homework. For this model, we used cross-validation to select the learning rate (0.1, 0.05 or 0.01) and the number of estimators (50, 75, 100, 150, 200 or 250) with the highest accuracy. To implement the model, we used the *GradientBoostingRegressor* from the *sklearn* library.

4.3.3 Random Forest

We also used a Random Forest model, another nonlinear ensemble method based on decision trees, to diversify the type of models that we used. As for Gradient Boosting, we used cross-validation to select the number of trees in the forest (50, 75, 100, 150, 200 or 250). To implement the model, we used the *RandomForestRegressor* from the *sklearn* library.

4.4 Feature Selection

As in the paper, we performed univariate feature selection using two methods: ANOVA and χ^2. The number of features selected was optimized as a model parameter with 5-fold cross-validation among the values (5, 10, 20, 30 or 50).

5 Results

5.1 Impact of Support

As we can observe in Fig.1, there is a difference in the amount of supporting actions between the friendships that ended in betrayal and those that did not. There is less support in the former ones. When looking at the imbalance between (potential) betrayers and victims, it seems higher in friendships that ended in betrayal. However, we should be careful because the standard errors are relatively high.

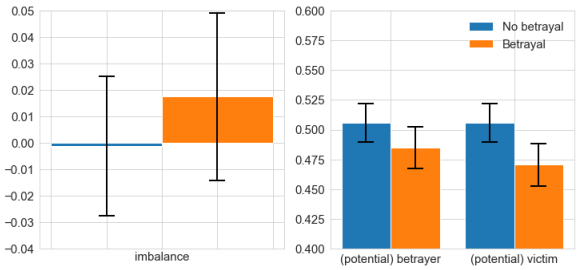

Figure 1: Support (average season score), imbalance plot on the left and mean value per player type on the right.

5.2 Analysis of Friendship Features

We plot in Fig.2 the correlation of the friendship features according to their types (mean or variance). We can observe that the means correlate with the Betrayal event, with a median of 0.04, while the variance correlates with a median of about 0.035. From this analysis it would seem that the variance provides a feature as valuable as the mean.

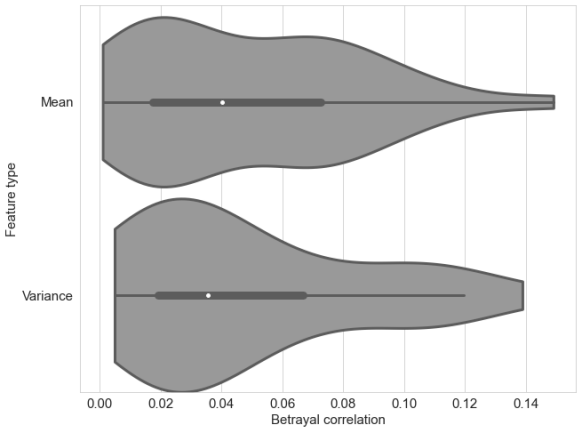

Figure 2: Violin plot of the feature types (mean or variance) correlations with the betrayal.

From the plot of the correlations of the player types (victim or betrayer) with the Betrayal event (Fig.3), we can observe that both features related to victims and betrayers are significant to predict betrayal. This corresponds to the statement made in the original paper.

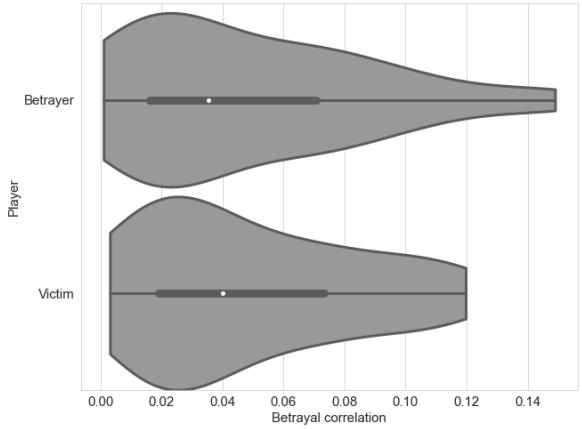

Figure 3: Violin plot of the player types (betrayer or victim) correlations with the betrayal.

5.3 Betrayal Prediction Results

In Table 1, we compare the results that we obtained by training the three models described above on two sets of features: a selection of those used in the paper and a selection of the friendship features created by us. As normalization of the features did not lead to a significant improvement in accuracy, we do not include the values achieved with normalization in the table. For better readability, we only include results from univariate feature selection using χ^2, which generally provides better results than ANOVA. More extensive results can be found in the Jupyter Notebook.

We can observe from Table 1 that the model type does not improve the accuracy scores significantly in every situation, but we can observe that the Gradient Boosting Regressor seems more relevant for the features of the paper. However, using means and variances of the features over the friendships improves the scores as compared to only using the features over the seasons as it is done in the paper. For example, the logistic regression model gives 52.97% accuracy with the paper's features, and 57.26% accuracy when using our new friendship features. It can be noted that we did not manage to reproduce the paper's result of 57% accuracy for logistic regression.

Model	Features	Accuracy ($\pm 95\%$ ci)
LR	Paper	$52.97\% \pm 1.52\%$
	Friendship	$57.26\% \pm 2.30\%$
GBR	Paper	$56.51\% \pm 3.52\%$
	Friendship	$59.15\% \pm 4.14\%$
RF	Paper	$54.26\% \pm 0.84\%$
	Friendship	$60.07\% \pm 4.52\%$

Table 1: Accuracy of the different models, using $\chi 2$ for feature selection and without normalization with bootstrapped confidence intervals. (LR = Logistic Regression, GBR = Gradient Boosting Regressor, RF = Random Forest)

6 Conclusion

We can confirm, based on the accuracy rates we achieved, the general conclusion that the authors of the paper came to: "the intention to betray can leak through the language one uses". Regarding our own extensions, we found that parsing the dataset by friendship rather than by season has its advantages. This way of treating the data gave us access to a new feature: the support between the players within a friendship. We found that this feature does indeed provide information about the destiny of a relationship. Furthermore, when using friendships as a basis for approaching the data, it is interesting to note that the variances as well as the means are very useful features. Finally, in the beginning of this project one of the things that we set out to explore was whether the use of more powerful machine learning models could improve the prediction results achieved. Our work would lead us to say that, using the features from the paper, a Gradient Boosting Regressor model achieves better results than Logistic Regression. However, our best result was of 60.07% accuracy, using the additional features that we engineered and a Random Forest model. This lets us think that the choice of the model is of less significant importance than the choice of the features when trying to improve the accuracy of the predictions.

References

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber and Cristian Danescu-Niculescu-Mizil 2015. *Linguistic harbingers of betrayal: A case study on an online strategy game*. *Proceedings of ACL*, 1:1650–1659.