Analysing the Upworthy Research Archive

Team 3

Citadel European Data Open 2021

Influence of Headline Sentiment

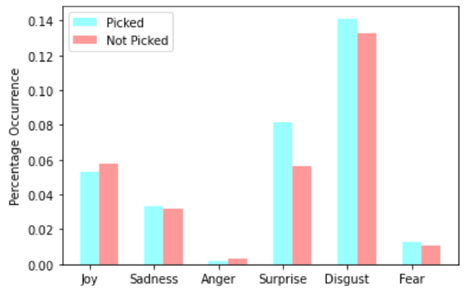
Sentiments of successful headlines tend to be negative:

In general, the picked headlines were more likely to have negative sentiments, whereas most non-picked headlines were mostly neutral. The majority of headlines were negative however, so this does not shed much light on why particular headlines were picked.

Sentiment	Headline Picked	Not Picked
Negative	44.0%	39.8%
Neutral	39.8%	42.8%
Positive	16.2%	17.5%

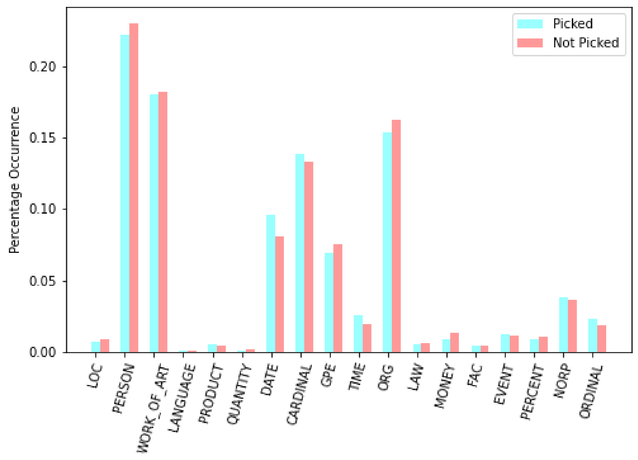
Emotions of successful headlines tend to be more sensational:

More insight can be drawn by looking at the emotions that headlines are associated to, as the headlines have more spread over emotions than over sentiments. The picked headlines tended to be more sensationalised and inflammatory, which is in-line with general criticisms of clickbait-like headlines. However, this analysis shows that moral or not, these type of headlines tend to achieve the best results.



Including Named Entities

Some patterns can be observed by including named entities. For example, picked headlines do tend to have more mention of numbers (cardinals and ordinals) than non-picked ones. Additionally, and perhaps more surprisingly, picked headlines are less likely to mention persons or organizations.



Questions of Interest

1. What can the components of successful headlines say about users' attitudes?
2. How do the headlines sentiments and click-through-rates (CTRs) vary with time?
3. Can headlines alone be used as a predictor for the CTRs?

Time Series Modelling

Time Series modelling techniques were used to model changes in CTR over the 2.5 year period. Residual heteroskedasticity is identified after fitting an ARIMA(0,1,1), using McLeod-Li hypothesis test (p-value obtained: 0.0036). This can be spotted visually be examining the 'pulsing' behavior of squared residuals of the model (Figure 1).

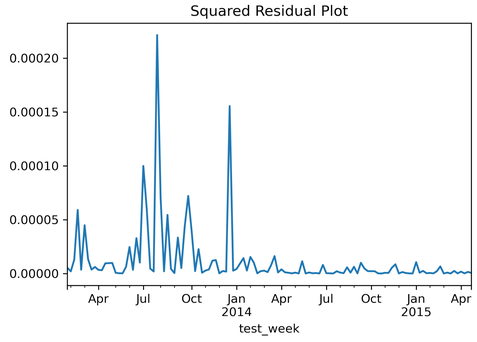


Figure 1:Residual Volatility Clustering

Accounting for this, the fitted model for CTR ($\{Y_t\}$) is an ARIMA(0,1,1)-GARCH(1,1):

$$Y_t = -0.000137 + Y_{t-1} + \epsilon_t + 0.486\epsilon_{t-1} + \eta_t$$
$$\eta_t = \epsilon_t(0.0000000171 + 0.13\eta^2_{t-1} + 0.866\delta^2_{t-1|t-2})$$

All model assumptions are found to be satisfied, but residuals are identified as non-normal.

Analysing the sentiment of picked packages over time showed that the proportions of negative, neutral and positive sentiment remain fairly stable. The proportion of positive sentiment over time could be modelled using an ARMA(1,2) model, the neutral and negative sentiment proportions are mostly white noise processes. It can be concluded, that although negative sentiment in the headlines lead to higher CRT, the proportion of negative headlines that outperformed neutral or positive headlines in causal statistical tests, is independent of time.

CTR Prediction

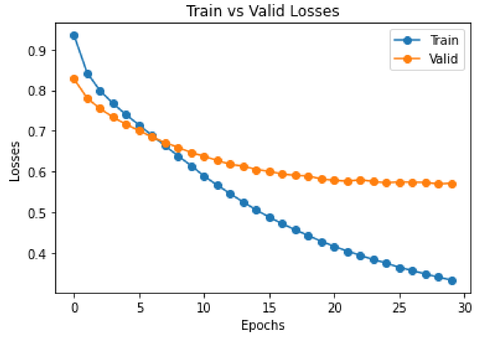


Figure 2:Learning Curves CTR prediction

Figure 2 shows the learning curves of our CTR prediction model, with full architecture described in our report. This model takes as input the headline BERT-based embeddings, publication time, headline's length, inferred sentiment scores, and mentions of named entities if any. It can serve to evaluate packages before deployment.

Winning Prediction

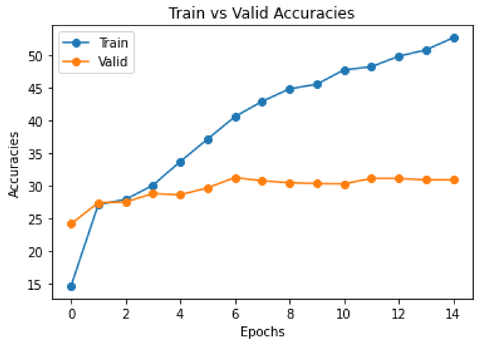


Figure 3:Winning Prediction: Accuracy

Figure 3 show the evolution of the training and validation accuracies of our Winning prediction model, with full architecture described in our report. We note that the validation accuracy remains quite low, with a maximum value of 30%, suggesting that for this simple model, the knowledge of the headlines of package does not suffices to predict the Winner of a given test.