CISA 4358 Final Report

4/28/2021

John Paul A Legreid

1. Problem definition

Toxic information continues to proliferate across the world wide web, and most platforms today use a combination of machine learning and manual review to keep the propagation of undesirable content to a minimum. Manual review poses problems to both the reviewer and the employer. Reviewers have been known to suffer PTSD-like symptoms from the sheer volume of horrible content they are required to review on a daily basis (Newton, 2020). Employers face more sanitary concerns of liability, scalability, and speed of content removal. We aim to solve these problems by applying machine learning techniques to achieve greater accuracy in classifying insincere questions on Quora. Our approach is to apply natural language processing techniques and supervised machine learning algorithms to solve this classification problem.

2. Data extraction

The data used was the Quora Insincere Questions training dataset, downloaded from Kaggle (n.d.). Data variables are shown below. The dataset contained 1,306,122 rows.

Variable	Definition	Description
qid	Question ID	A UUID-like string, perhaps the hash of the question text
question_text	Quora Question	A string containing the text of the question to be classified
target	Target	0 or 1 indicating sincerity of question (1 = insincere)

3. Data preparation

For this project, only 2,000 rows were selected, with equal amounts of sincere and insincere questions being selected. Since this is a dataset created by Quora labelers, there are no missing values and missing value imputation is not required. The quid column provided no useful information for classification and was consequently dropped. The `question_text` column was vectorized and converted to bag of words using n-grams 1-3. Stop words were removed using a standard list of English stop words.

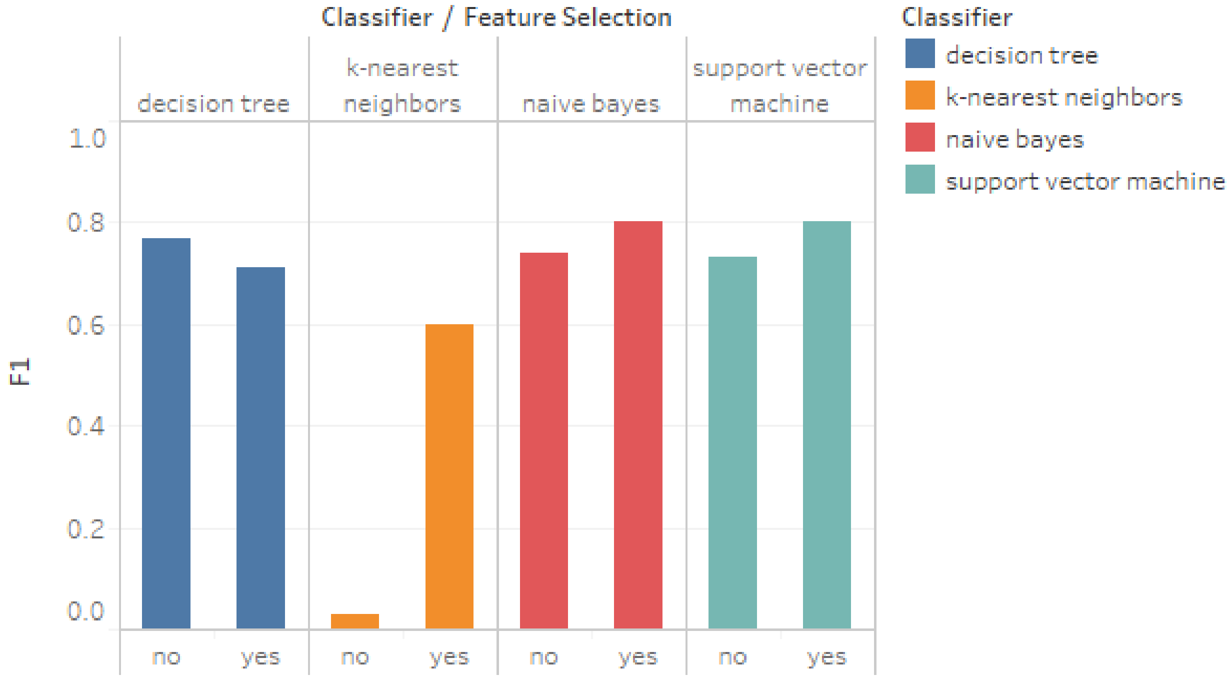
4. Predictive modeling

Chi-squared feature selection was performed, and features with $p < 0.05$ were chosen. Machine learning algorithms naïve bayes, decision tree, k-nearest neighbors, and support vector machine were applied to the data. Each model was trained with and without feature selection. The results, sorted by descending F1 score, are presented in Table 1 below. Figure 1 presents the results in a more easily digestible graphical format. Please see the accompanying Jupyter Notebook for additional information.

Table 1

classifier	feature	accuracy	precision	recall	f1
naive bayes	yes	0.79	0.79	0.80	0.80
support vector machine	yes	0.80	0.76	0.84	0.80
decision tree	no	0.77	0.73	0.81	0.76
naive bayes	no	0.74	0.71	0.77	0.74
support vector machine	no	0.73	0.72	0.75	0.73
decision tree	yes	0.74	0.61	0.85	0.71
k-nearest neighbors	yes	0.68	0.47	0.84	0.60
k-nearest neighbors	no	0.49	0.01	1.00	0.03

Figure 1



F1 as an attribute for each Feature Selection broken down by Classifier.
Color shows details about Classifier.

References

Kaggle. (n.d.). Quora Insincere Questions Classification. [train.csv]. Retrieved from https://www.kaggle.com/c/quora-insincere-questions-classification/data?select=train.csv

Newton, C. (2020, May 12). *Facebook will pay $52 million in settlement with moderators who developed PTSD on the job.* Retrieved from https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health