Predicting reddit score using word2vec model

Laxman Panthi[1,2]

[1] Harrisburg University of Science and Technology

[2] Medical Mutual of Ohio

Author Note

MS in Analytics, Department of Analytics, Harrisburg University

Data Engineer, Medical Mutual of Ohio

Correspondence concerning this article should be addressed to Laxman Panthi, Cleveland, Ohio 44130. E-mail: lpanthi@my.harrisburgu.edu

Predicting reddit score using word2vec model

```
## unable to import 'smart_open.gcs', disabling that module
```

Introduction

Text data analytics has been an interesting conundrum in the area of analytics overall. Several methods have been utilized by experts over time and cutting edge algorithms coming in everyday to accurately depict the human level interpretation of the text data. There are different data sources available in the internet to analyse these kind of model with the top ones being social media platforms and forums like Twitter and Reddit.

Reddit is a good source of information and opinion from people around the world. A study was done previously to predict faruadulent transactions happening over reddit. Unlike any other marketplace that has a proper rating system for buyers to decide, Reddit which is not a marketplace does not have a user rating system. However, transactions do happen on Reddit. Reddit neither keeps detailed personal information nor keeps a transaction history of its users. This is not helpful for its members when they want to make a transaction with another user. Several fraudulent transactions happen every day on Reddit. There is no proper way of differentiating legitimate sellers from fraud (Landstein, 2020). The data collected about the user such as the age of account, Karma (upvotes), Verified Email Address, Gold, Comments, Moderator, Subreddits visited and Trophies were used to develop a model using Multiple Logistic Regression to predict if the user was Redditor or Scammer based on the Reddit's categorization of these users. Redditors were identified from the transaction previously completed successfully and Scammer identified from the Reddit's banned list.

Methods

Texts can be represented into vector form to be able to use in the predictive models. As studied by the vector representation of the words and creating word similarity (Mikolov, Chen, Corrado, & Dean, 2013), it is possible to find linkage between different words in a sequence. The word 2 vec model takes into consideration the sequential nature of high frequency of texts and creates model to represent those words as vectors(Mikolov et al., 2013). A model can be developed to use similar type of words in a prediction algorithm and utilized to performed advanced level of text analysis (Pagolu, Reddy, Panda, & Majhi, 2016). Logistic regression is then applied to the vectors such generated.

Problem Statement

Although the people of internet are usually cautious about posts that are not safe for work, it is essential for Reddit to identify the records that are adult rated. We are solving a similar type of problem in this project. The dependent variable is the field called over_18 that signifies if the post is adult rated or not and we are looking to predict that using the text in the title field.

Dataset

The data was downloaded from kaggle (Fontes, 2020). It conmtains all the posts related to COVID-19 pandemic in the r/dataisbeautiful subreddit of reddit.

Data Description

- id - Unique identifier
- title - Title of the reddit post
- score - Reddit score

- author - Author of the post
- author_flair_text - Author's flair
- removed_by - Who removed the post?
- total_awards_received - Total number of awards
- awarders - Total number of awarders
- created_utc - Created at
- full_link - Link of post
- num_comments - Number of comments
- over_18 - True if not safe for work (nsfw)

Data Exploration

title
data_irl
Police killing rates in G7 members [OC]
What's getting cut in Trump's budget
Almost all men are stronger than almost all women [OC]
America's new tobacco crisis: The rich stopped smoking, the poor didn't
Rolls of toilet paper used per person per year [OC]
Tinder over 3 years (18-21 Male) [OC]
United States of Apathy: 2016 US Presidential Election Results if Abstention from Voting Was Co
I made a chart showing the popular vote turnout in 2008, 2012 and 2016. Hillary didn't lose becau
Brexit: London, Scotland, N. Ireland, and younger generations voted to remain. Almost everyone e

Results and Discussion

```
## /Users/laxmanpanthi/anaconda3/envs/anly540/lib/python3.7/site-packages/gensim/models/
##   "C extension not loaded, training will be slow. "
```

```
## accuracy 0.9872717210846708
```

```
##               precision    recall  f1-score   support
##
##         TRUE       0.99      1.00      0.99      1784
##        FALSE       0.00      0.00      0.00        23
##
##     accuracy                           0.99      1807
##    macro avg       0.49      0.50      0.50      1807
## weighted avg       0.97      0.99      0.98      1807
##
##
## /Users/laxmanpanthi/anaconda3/envs/anly540/lib/python3.7/site-packages/sklearn/metric
##   _warn_prf(average, modifier, msg_start, len(result))
```

The accuracy is 98% which indicates that the over_18 field can be predicted using the text of the title of the post. Further work can be done in this model to include all the variables that were part of the original dataset.

R (Version 3.5.1; R Core Team, 2018) and the R-package *papaja* (Version 0.1.0.9942; Aust & Barth, 2020) has been used extensively for this study for data analysis and the creation of this report.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Fontes, R. (2020). Reddit - data is beautiful. *Kaggle.* Retrieved from https://www.kaggle.com/unanimad/dataisbeautiful/version/4

Landstein, E. (2020). How to use machine learning to make prediction on reddit: Multiple logistic regression. *Medium.* Towards Data Science. Retrieved from https://towardsdatascience.com/how-to-use-machine-learning-to-make-predictions-on-reddit-part-i-44cd210ec427

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781.*

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (scopes)* (pp. 1345–1350). IEEE.

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/