Capstone Project: Proposal

Flávio Henrique de Freitas
(Dated: August 16, 2018)

I. DOMAIN BACKGROUND

Social networks websites have become an important communication tool and source of information. The hours spent in average connected per day in the past years is up to 6 hours [1] for adults and 9 for teenagers, while 30% of this time is on social networks [2]. During a normal navigation on such platforms, users are exposed to several posts such as friends' statuses, images, news and more. With such amount of information and variety of content, the time for the user to decide to interact with the content is very small. Gitte at al. [3] suggest that we take around 50 milliseconds to make a good first impression and this has proved to be very powerful in a wide range of contexts.

Besides being a place for connecting with friends and sharing moments of the user's life, a survey has shown that social networks are also used as a source of news and information by 67% of the users [4]. Part of these posts are articles that can be read on an external website. Typically such posts show the title of the article and sometimes a small part of its content and an image.

Considering the offer of content and competition with so many interesting posts, showing a proper title for the post affects the probability that a user will check the content. This measure has a strong impact on how many readers an article will have and how much of the content will be read. Furthermore, showing the user a content they prefer (to interact) increases the user satisfaction. It is thus important to accurately estimate the interaction rate of articles based on its title.

II. PROBLEM STATEMENT

When an author writes a text, it is expected that their words will influence and bring value to the readers. While writing, the title is one of the important details that needs to be taken in consideration, because this will normally be the first contact place of their work. Thus, to create a good first impression, to have more people read the article and interact with it, choosing a good title is very important.

Some of the most used platforms to spread ideas nowadays are Twitter [5] and Medium [6]. On the first one, articles are normally posted including external URLs and the title, where users can access and demonstrate satisfaction with Favorites or Retweeting (sharing) of the original post. The second one shows the full text with tags to classify the article and "Applause" (similar to Twitter's Favorites) to show how much the users appreciate the content. A correlation between these two networks can bring us more valuable information.

The problem to be solved: *Predict the range of favorites and shares count an article receives based on its title; and analyze how the title length and the tags have performed.*

III. DATASETS AND INPUTS

The data used to predict how titles will perform was gathered from the accounts of the non-profit organization FreeCodeCamp on Medium [7] and Twitter [8]. With both social platforms, it was possible to get public information about how the users interacted with the content, such as Favorites and Retweets from Twitter, and "Applause" from Medium.

The reason to correlate the number of Favorites and Retweets from Twitter with a Medium article, was to try to isolate the effect of number of reached readers and number of Medium Applauses. Because the more the article is shared in different platforms, the more readers it will reach and the more Medium Applauses it will receive. Using only the Twitter statistic, it is expected that the articles reached initially almost the same number of readers (that are the followers of the FreeCodeCamp account on Twitter), and the performance and interactions with it are limited to the characteristics of the tweet, for example, the title of the article, that is exactly what we want to measure.

It was decided to choose FreeCodeCamp account, because the idea is to limit the scope of the subject of the articles and predict better the response on a specif field. The same title can perform well on one category (e.g. Technology), but not necessarily on a different one (e.g. Culinary). Also this account posts as the Tweet content the title of the original article and the URL on Medium.

After getting the articles from FreeCodeCamp written on Medium and shared on Twitter, there is a dataset of 719 data points. Here are some examples of such correlation:

IV. SOLUTION STATEMENT

Classification is a common task of machine learning (ML), which involves predicting a target variable taking in consideration the previous data [9]. To reach such classification, it is necessary to create a model with the previous training data, and then use it to predict the value of the test data [10]. This process is called Supervised Learning, since the data processing phase is guided toward the class variable while building the model [11].

TABLE I. Sample of the data points

Title	Retweet	Favorite	Applause
ES9: JavaScript's state of the art in 2018	15	48	618
Here's another way to think about state: How to visually design state in JavaScript	10	30	2
How to understand Gradient Descent, the most popular ML algorithm	4	14	102

TABLE II. Complete description of the dataset fields

Field	Description
Title	The content of the tweet, FreeCodeCamp normally uses the title of the article from Medium and sometimes the username of the author from Twitter
Retweet Count	How many times that tweet was "Retweeted" on Twitter
Favorite Count	How many times that tweet was marked as favorite on Twitter
Medium Applauses	How many times that article was marked as favorite on Medium
Medium Categories	Which tags were used to tag the article on Medium
Created at	When the tweet was posted
URL	The website of the article on Medium

Predicting number of shares and favorites of an article can be treated as a classification problem, because the output will be discrete values (range of shares and favorites). As input, the title of the articles with each word as a token t1, t2, t3, ... tn.

For this task we will evaluate the following algorithms: Support Vector Machines (SVM), Decision Trees, Gaussian Naive Bayes (GaussianNB), K-Nearest Neighbors and Logistic Regression. In the end, it will be compared the performance of each one of them and one will be chosen. To estimate accuracy, it will be used a 5-fold cross validation, that splits the dataset in 5 parts, 4 of trainning and 1 of testing. The implementation of this project will be made usying Python, Numpy [12] and Scikit [13].

V. BENCHMARK MODEL

This project will run the same testing and training data for multiple algorithms, the comparison between them can be used to evaluate the overall performance.

VI. EVALUATION METRICS

At least one evaluation metric is necessary to quantify the performance of the benchmarks and solution model. For this project, it will be used the accuracy, which is the number of correct predictions made as a ratio of all predictions made.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions made}}$$

This metric only works well if there are similar number of samples belonging to each class. For this reason, we will divide the range of shares and favorites count in a way that respects this distribution.

VII. PROJECT DESIGN

The steps to solve the proposed problem will take as reference the one proposed by S. Raschka [14]. They are:

1. **Data gathering**: This step is responsible to get the datasets that will be used on to analyze, train and test the models. This data was already gathered from Twitter and Medium and aggregated like showed in the section "Datasets and Inputs".
2. **Data pre-processing:** The dataset will be cleaned, formatted or added the missing values.
3. **Exploring Data**
   (a) **Prepare environment to run the simulations:** The environment used to make this simulation will be a Jupyter Notebook. For each of the steps, we will describe what is expected and show the Python code used for the implementation.
   (b) **Training and Testing Data Split:** It will be defined the sets for training and testing. From the overall data points that we have 719, 143 will be testing data and 576 training.
4. Training and Evaluating Models
   (a) **Model Performance Metrics Implementation:** The chosen Evaluation Metric will be implemented to analyze how the each of the models performed.
   (b) **Models Implementation:** Implement all the algorithms of supervised learning chosen to test how each of them perform for such dataset.
   (c) **Model Performance Metrics:** Evaluate all the models and make a comparison between their performance.
   (d) **Model Tuning:** Tune the algorithms to try to find the best parameters.

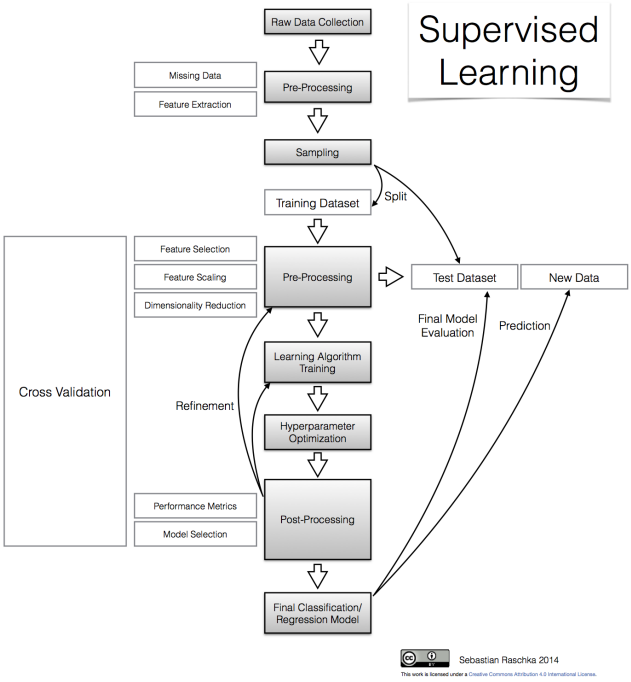

FIG. 1. Sebastians Raschka's workflow [14]

(e) **Reiterate:** Reiterate the previous steps and check how the performance is evolving.

5. **Choosing the Best Model:** Decide the best model to make the desired prediction.

REFERENCES

[1] *eMarketer Report. (2017). US Time Spent with Media: eMarketer's Updated Estimates for 2017. Accessed 9 Aug. 2018. Available at: https://www.emarketer.com/Report/US-Time-Spent-with-Media-eMarketers-Updated-Estimates-2017/2002142.*

[2] *Common Sense Media. (2015). The Common Sense Census: Media Use by Tweens and Teens. Accessed 9 Aug. 2018. Available at: https://www.commonsensemedia.org/research/the-common-sense-census-media-use-by-tweens-and-teens.*

[3] *Lindgaard, Gitte Fernandes, Gary Dudek, Cathy M. Brown, Judith. (2006). Attention web designers: You have 50 milliseconds to make a good first impression! Behaviour and Information Technology, 25(2), 115-126. Behaviour IT. 25. 115-126. 10.1080/01449290500330448.*

[4] *Shearer, E., Gottfried, J. (2017). News use across social media platforms 2017. Accessed 9 Aug. 2018. Available at: http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/.*

[5] *Twitter. Accessed 13 Aug. 2018. Available at: https://www.twitter.com.*

[6] *Medium. Accessed 13 Aug. 2018. Available at: https://www.medium.com.*

[7] *freecodecamp. Accessed 13 Aug. 2018. Available at: https://medium.freecodecamp.org/.*

[8] *freecodecamp.org. Accessed 13 Aug. 2018. Available at: https://twitter.com/freecodecamp.*

[9] *N. Abdelhamid, A. Ayesh, F. Thabtah, S. Ahmadi, W. Hadi. MAC: A multiclass associative classification algorithm J. Info. Know. Mgmt. (JIKM), 11 (2) (2012), pp. 125001-1-1250011-10 WorldScinet.*

[10] *I.H. Witten, E. Frank, M.A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, MA (2011).*

[11] *Thabtah, S. Hammoud, H. Abdeljaber, Parallel associative classification data mining frameworks based mapreduce, To Appear in Journal of Parallel Processing Letter, March 2015, World Scientific, 2015.*

[12] *NumPy. Accessed 14 Aug. 2018. Available at: http://www.numpy.org/.*

[13] *scikit-learn. Accessed 14 Aug. 2018. Available at: http://scikit-learn.org/stable/.*

[14] *S. Raschka. Predictive modeling, supervised machine learning, and pattern classification. Accessed 14 Aug. 2018. Available at: https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html.*