Implementation of The Random Forest Method on The Retweet Classification Model Based on Content

1st Akmal Ariq Santoso
Faculty of Informatics
Telkom University
Bandung, Indonesia
akmalariq@students.telkomuniversity.ac.id

2nd Jondri
Faculty of Informatics
Telkom University
Bandung, Indonesia
jondri@telkomuniversity.ac.id

3rd Kemas Muslim Lhaksmana
Faculty of Informatics
Telkom University
Bandung, Indonesia
kemasmuslim@telkomuniversity.ac.id

Abstract

Twitter as one of the biggest social media on the internet has been used as the center of information exchangeon mainstream media. As this paper was written Covid-19 information sporadically propagated through twitter. To help spread validated information to the masses we need to understand which factors are relevant and support the information diffusion. In this paper author tried to find similarities between tweetsby using TF-IDF, author also applied content features from tweet's meta-data to random forests classifier to predict which tweets users might retweet. The result of the shows that by using content features, machinelearning models can predict retweets from users. The proposed method of combining content features fromtwitter metadata and TF-IDF leads to a better model than the stand-alone features with 69.97% of accuracy.

Keywords: Retweet Prediction, TF-IDF, Random Forests

I. INTRODUCTION

Twitter, a micro blogging media has been a center of information exchange people used widely on the internet. Twitter is a follow-based system, user A can see tweets from user B in user A's home timeline by following user B. Twitter users can "comment", "like", and "retweet" a tweet they want to interact with. Retweetis used to share or repost tweets other users' tweet as yours, this can result in your followers see tweets from usersthey don't follow. This sharing of tweet or retweet can cause a spread of information across twitter users as described in [1].

Information diffusion studies[2] described twitter as a powerful tool for rapid information delivery in emergency. According to [2], social media such as twitter can also shape social consciousness, that is why in thisstudy we tried to better understand which factors are best correlated to user's retweet. Study [3] used Covid-19 related tweets as dataset on their experiment, in this paper we tried vaccine related tweets. The study [4] shows random forest has the best prediction accuracy between other models. Therefore, we had conducted a study of retweet prediction from vaccine related tweets or tweets that contain the word "vaksin" with content-based features and TF-IDF on random forests classifier.

II. LITERATURE REVIEW

We have divided the problems on retweet prediction into three parts[5], namely 1) Which tweet will the userretweet? 2) Who is the target user who will retweet a tweet? 3) Why does one tweet get more retweets than another? Therefore, we had focused on the first problem. We tried to predict which tweets are more prone to be retweeted publicly by other users. Then we divided the features based on twitter data attributes into three groups[6] 1) User- based 2) Time-based 3) Content-based. The result of [6] showed that user-based features especially the number offollowers and following are highly correlated to retweet count in their models. From the result, we wanted to focusmore on content-based features since we want to understand what features are best used for tweets from users with similar follower counts. Study [7], tried TF-IDF and DistilBERT as feature extraction methods to find similarities between one tweet and another. In which the result of the study came as equal between DistilBERT and TF-IDF, though since the TF-IDF method is more straightforward to implement we went to use TF-IDF for a faster outcome. Another study [8] showed methods using TF-IDF and Doc2Vec as text-retrieval methods then used the result as input feature to various machine learning methods, with TF-IDF as the better tool than Doc2Vec.Since to utilize Doc2Vec the researchers[8] believed it may need thousands of words in each document (tweet) forbetter result, in which a twitter has a limit of 280 characters for each tweet. Therefore, we believe it is better to useTF-IDF in our study. We tried to combine content-based aspect of tweets and TF-IDF as features on random forests.

III. METHOD OF RESEARCH

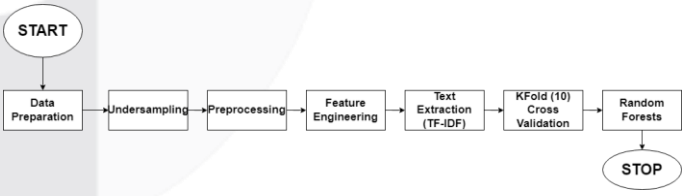

Figure 1. System Flowchart

The proposed approach flow chart can be seen in figure 1.

A. Data Preparation

Data crawled with the keyword "vaksin" from Tweepy from 13th July-4th August 2021 has been collected with 128.741 tweets. Since retweet count had a high correlation with follower count[6], to avoid bias to accounts with different range of followers we only took tweets from accounts with less than 10.000 followers resulting in 119.721 tweets.

B. Under sampling

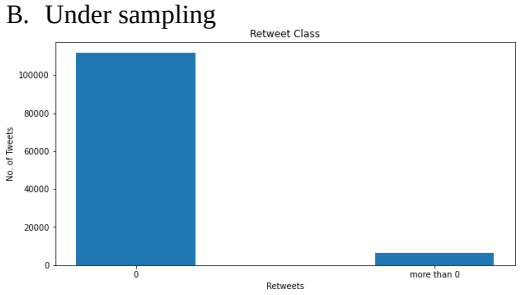

Figure 2. Imbalance Retweet Class Distribution

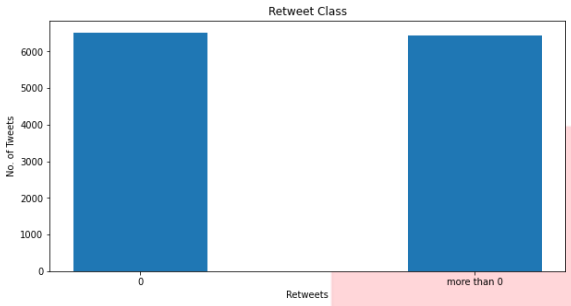

Figure 3. Balance Under sampling Retweet Class Distribution

Random under sampling is done to give a fair advantage between each target class since as shown in figure 2the data is imbalance with unretweeted has 111.794 tweets and retweeted has 6.555 tweets. We randomly sampled the unretweeted to 6.555 tweets to give a fair advantage between each class.

C. Preprocessing

The prepared data was checked for duplicates. Then the data was checked for tweets that doesn't containthe keyword "vaksin". As shown in study [9] the tweet's text contents are lower-cased and stripped from 1) Strings that are less than 3 characters 2) Punctuations 3) Number characters.

D. Feature Engineering

The data then added 4 new features to include content-based features:

- *has_media* {0,1}
- *optimum_text_length* {0,1}
- *has_url* {0,1}
- *has_hashtag* {0,1}

For the target the retweets are grouped to 2 classes, tweets with 0 retweet count and tweets with retweet countmore than 0.

E. Text Extraction

To include text content from the tweets text extraction (TF-IDF) is implemented.

Term frequency: $tf(t,d) = \frac{f\ t,d}{\sum tFgft,d}$ (1)

Inverse document frequency:

$$idf(t,D) = log\ \frac{N}{|\{d \in D : R \in d\}|} \quad (2)$$

The result of TF-IDF is stored in a sparse-matrix form.

To avoid loss in information of TF-IDF we used column transformer to use different methods on different feature columns of the data.

F. K-Fold

Cross-validation is a method of resampling data to ensure the accuracy of the model is unbiased. Since the data we have is limited to 13.110 tweets, cross-validation is chosen as the best method to give precise accuracy tounseen data. The K-Fold method will shuffle the data and resample it to 10 train and test datasets. These datasets then will be used in the random forests classifier in which the mean accuracy scores of the datasets is taken as thegeneral result of the models.

G. Random Forest

Random Forests is an ensemble learning method, as decision trees have a tendency to overfit on training sets. Random forests act as a collection of decision trees of different depths from its bootstrap sampling method which result in creating different tree models from one training dataset. In this study we utilize random forests library from the scikit-learn library [10].

IV. RESULT RESEARCH AND DISCUSSION

A. Result

Gambar dinomori secara berurutan. Letak penulisannya di bawah gambar yang dijelaskan. Contoh: Gambar 1(A)

Table 1. Results

Features	Accuracy Macro	Precision Macro	RecallMacro	F1-Score Macro
Content-based	68.24%	69.26%	68.17%	67.69%
TF-IDF	66.97%	67.00%	66.95%	66.94%
Content-based + TF-IDF	**69.97%**	**69.99%**	**69.97%**	**69.97%**

Accuracy shows the number of correctly classified tweets divided by the total number of tweets as shown inEq. 3. Other than accuracy we have used other metrics such as precision to show the model's performance in predicting true positive outcomes out of all the positive (retweeted) instances and recall showing true positive outcomes out of all correct results (true positive and false negative). To show the balance on precision and recall we also add F1-score as the average rate of the model's performance.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \quad (3)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

$$F1\ Score = 2x\ \frac{Precision + Recalll}{Precision + Recall} \quad (6)$$

B. Discussion

The result in this study is not significant, although it might have a slight relation between the two prominent features, we have concluded there are features we can add. In future study, it might be better to use more content- based features such as [3] which used 20 content-based features and add features as mentioned in [11] which focused on emotion and topic. We

also would like to assess that binary classification might show a better result than multiclass classification.

1.

V. CONCLUSION

We can conclude from the study that a content similarity of tweet text content with TF-IDF can be used to predict retweets from users. Content-based features mentioned in this study can be used for retweet prediction. Although the result of this study is not significant, from the slight improvement from our proposed model we canconclude that we are on the right approach for finding features that influence users to propagate information.

REFERENCE

[1] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks:Models and methods," *Inf.*, vol. 8, no. 4, 2017, doi: 10.3390/info8040118.

[2] J. Kim, J. Bae, and M. Hastak, "Emergency information diffusion on online social media during stormCindy in U.S.," *Int. J. Inf. Manage.*, vol. 40, no. February, pp. 153–165, 2018, doi: 10.1016/j.ijinfomgt.2018.02.003.

[3] G. Piao and W. Huang, "Regression-enhanced random forests with personalized patching for covid-19 retweet prediction," *CEUR Workshop Proc.*, vol. 2881, pp. 13–16, 2020.

[4] H. Bunyamin and T. Tunys, "A comparison of retweet prediction approaches: The superiority of random forest learning method," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 14, no. 3, pp. 1052–1058, 2016, doi: 10.12928/TELKOMNIKA.v14i3.3150.

[5] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet: A popular information diffusion mechanism – A survey paper," *Online Soc. Networks Media*, vol. 6, pp. 26–40, 2018, doi: 10.1016/j.osnem.2018.04.001.

[6] T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter – Analysis of predictivefeatures," *J. Comput. Sci.*, vol. 28, pp. 257–264, 2018, doi: 10.1016/j.jocs.2017.10.010.

[7] N. Shoeibi, N. Shoeibi, P. Chamoso, Z. Alizadehsani, and J. M. Corchado, "Similarity approximation ofTwitter Profiles," no. June, pp. 1–18, 2021, doi: 10.20944/preprints202106.0196.v2.

[8] I. Daga, A. Gupta, R. Vardhan, and P. Mukherjee, "Prediction of likes and retweets using text information retrieval," *Procedia Comput. Sci.*, vol. 168, pp. 123–128, 2020, doi: 10.1016/j.procs.2020.02.273.

[9] F. Resyanto, Y. Sibaroni, and A. Romadhony, "Choosing The Most Optimum Text Preprocessing Method for Sentiment Analysis: Case:iPhone Tweets," *Proc. 2019 4th Int. Conf. Informatics Comput.ICIC 2019*, pp. 2–6, 2019, doi: 10.1109/ICIC47613.2019.8985943.

[10] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. May2014, pp. 2825–2830, 2011.

[11] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet Prediction based on Topic, Emotion and Personality," *Online Soc. Networks Media*, vol. 25, no. August, p. 100165, 2021, doi: 10.1016/j.osnem.2021.100165.