Predicting Information Diffusion in Online Social Platforms: A Twitter case study.

Kateryna Lytvyniuk, Rajesh Sharma, and Anna Jurek-Loughrey

Abstract Online social media has become a part of everyday life of modern society. A lot of information is created on these platforms and shared with the community continuously. Predicting information diffusion on online social platforms has been studied in the past by many researchers as it has its applications in various domains such as viral marketing, news propagation etc. Some information spreads faster compared to others depending on topic of interest of the online users. In this work, we investigate the information diffusion problem with Twitter data as a use case study. We define tweet popularity as number of retweets any original message receives. In total we extracted 27 features which can be categorised into content, user, sentiment and initial retweeting behaviour for creating our prediction model. We study the problem of predicting as a multiclass prediction task. Three datasets from Twitter about three different topics are collected and analysed for building and testing various models based on different machine learning algorithms. The models were able to predict up to 60% of overall accuracy and an F1 score of 67% is obtained. The models were created using one of the dataset and tested on all the datasets, which shows that the model is robust enough to handle information diffusions associated with different topics.

1 Introduction

Social media platforms allow Internet users to create and consume content in a very convenient and quick way. The influence of such online networks is very high as the Internet has become the primary source of new information in the present society. Understanding the information diffusion processes on these networks may help addressing many real world challenges such as investigation and prevention of terrorism activity [4]. The data on social media can also be analysed with an objective of observing the trends of elections results [1], correlating events between social media platforms such as Twitter and stock market [2].

Kateryna Lytvyniuk, Rajesh Sharma
University of Tartu, Estonia, e-mail: kateryna.lytvyniuk@ut.ee, rajesh.sharma@ut.ee,

Anna Jurek-Loughrey
Queen's University Belfast, UK e-mail: a.jurek@qub.ac.uk

In this work, we investigate the information diffusion problem using Twitter, which is one of the most popular social platform used by internet users. According to recent updates there are 335 million active users on Twitter[1]. Messages (referred to as Tweets) appear continuously and spread according to the interest of the Twitter's users. Each message can be forwarded by other users, also called as retweeting, or it can be liked or commented by others. These all activities helps in spreading the message through the network.

Various studies have been performed by researchers, that aimed to analyse information diffusion in online social platforms and to understand why certain messages are more popular than others [5, 9, 15]. People often express their opinions about specific topic or events which are most relevant to them or related to present real world events. Some studies have been very specific in investigating the diffusion of information with respect to news [14], advertising campaigns [10]. As opposed to previous research, our study is about predicting information diffusion and is topic independent. In particular, we investigate following research questions (RQ):

1. RQ 1: How can information diffusion be modelled?
2. RQ 2: What features are the most discriminative for the diffusion prediction?
3. RQ 3: How well a message diffusion can be predict using the identified features?
4. RQ 4: How initial retweet activity can help in predicting tweet popularity?
5. RQ 5: Is it possible to predict tweet popularity in a coming time window (for example an hour) based on tweet's behaviour in previous time window?

The first three research questions have been extensively studied in the past, however, RQ4 and RQ5 have not gained much attention from the research community.

We study these five research questions by first collecting dataset from Twitter about three topics namely, 1) *Cryptocurrency*, 2) *Smartphone brands*, 3) *Football*. We created our model using the tweet data associated with *Cryptocurrency*. We extracted in total 27 features which are either related to 1) the users who are tweeting , 2) the content of the tweets, 3) the sentiment associated with the tweets and 4) the initial behavior of the tweets, which basically calculated the number of times a tweets attracts the retweet in some initial time period. We study the importance of various features and then incorporate most relevant features into our models for predicting the retweeting behavior.

With respect to RQ 5, we are interested in the tweet diffusion for a particular tweet T_w. In other words, given the tweet diffusion pattern for a certain time period t, what would be the retweeting behavior for the tweet T_w in the next time window $t+\delta$. The problem in the past has been studied in different forms such as by calling it popularity [5] or interestingness [9]. In our study,

[1] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

for modelling the prediction problem, we used four different machine learning algorithms namely 1)Random Forest, 2) Tree Bag, 3) Gradient Boosting Machine and 4) Xgboost. We studied this problem as a multiclass classification problem inline with [5]. We created different sets of classes where each class represents a range of retweet numbers. We then predicted the class (or the range) for each tweet. The model created using the *Cryptocurrency* topic, was tested on the other two datasets that is *Smartphone brands* and *Football*. Using our methodology, we were able to predict up to 60% of overall accuracy and a F1 score of 67 was obtained.

The rest of the paper is organized as follows. In Sec. 2, we present related work. In Sec. 3, we describe our dataset and in Sec. 4 methodology is described. In Sec. 5 we discuss results of our research investigations. Sec. 6 concludes the work by also discussing the future works.

2 Related Work

In this section, we discuss relevant literature which are either closely related or have some overlaps with respect to our research. In [15], the authors investigated how topics spread through network structures in particularly analysing the impact of the presence of @username in tweets, which basically means involving users more proactively in the diffusion process. They concluded that a topic might have a different propagation efficiency at different time stages of its lifecycle.

A lot of previous work has focused on popularity prediction problem using the number of future retweet as a measure of the popularity. In [5] the authors tried to find out what different factors influence information propagation on Twitter. They focused on content features of the tweets as well as the user features who initiates the tweets. Since it is difficult to predict the exact number they put forward the concept of multiclass classification for the prediction task. In contrast to [5] in [9] the researchers predicted the retweeting behavior using tweet characteristics only. They discovered that there is not such a strong correlation between number of followers of a user initiating the tweet and retweet count. It was concluded that messages were more likely to be retweeted if they were about a general popular topic compared to a specific personal topic.

Most of the studies which have modelled information cascades have either focused on user properties, such as importance of the users [7], or on network properties such as edge growth rate, diameter and degree distribution [12]. Some other works have analysed the diffusion by following explicit paths of the propagation [6, 13]. In [6], after collecting data of tweets and users activities, Tweet-Trees are created, which represent tweet propagation in the network. Nodes of a tree represent followers of the user that has retweeted the initial message. Authors used each tweet's linguistic features and a profile of initial creator of the tweet to make the prediction. In a similar approach, in [13], authors analysed information cascades from one user to another from

a stream of tweets and the social graph. They not only focused on tweets but also considered other types of information over a social network e.g. links or hashtags. In addition, they suggested methods to deal with the missing data, with regards to constructing the information cascades.

In [**?**] the authors used the propagation of Twitter posts to model the influence of users. As oppose to the other papers, they considered reposting, rather than re-tweeting, to indicate the diffusion level. For every initial post an influence tree (cascade) was generated, where the number of users included in the tree defined the influence score of the seed post. Two types of features were used as predictors: user-related attributes (e.g. number of followers) and characteristics related to user's past performance. The authors also investigated the role of a post's content in the propagation process. According to the authors' findings, the average number of past reposts by users immediate followers and the number of user's followers can be used to predict future influence of a user (propagation of the user's posts). The content of a post was not found to improve the predictive performance.

This work is different from the application oriented work such as prediction of tweet popularity related to breaking news analysis [14], or study of advertising campaign strategy based on information cascades in Twitter [10], or study of brand popularity prediction in social networks [8]. Our work is more close to that of [3, 5, 9] however, compared to all the previous studies which have mainly considered user and content features, we also explored the sentiments present in the tweets and the initial propagation of the tweets in predicting the information diffusion.

3 Dataset

In this section, we describe the dataset we used for our analysis. We wrote a Python script for collecting data from Twitter, using Streaming API in JSON format. The collected data is related to three different topics: *Cryptocurrency*, *Smartphone brands* and *Football*. Description of the collected data is provided in the Table 1.

Table 1: Dataset description

Dataset	# tweets	# original tweets	Keywords	Duration	Train or Test	Description
Cryptocurrency	3,110,500	1,606,696	cryptocurrency,bitcoin, blockchain, etherium	Jan-Feb, 2018	Both	Tweets about cryptocurrency trends.
Smartphone brands	601,380	340,504	'Samsung', 'Huawei', 'Xiaomi', 'iPhone', 'Lenovo', Nokia, 'LG', 'smartphone'	April, 2018	Test	Tweets about some of most popular smartphones brands.
Football	192,593	103,755	football, World Cup	June, 2018	Test	Tweets about football and in particular 2018 FIFA World Cup.

Each Twitter message is represented in JSON format and consist of many attributes. Besides the message content, metadata of tweet is extracted as well. This data includes user profile, location, statuses, counts of entities (such as special symbols, links) and language. User data consists of user profile characteristics. The most relevant for our research are: number of followers,

number of friends and date of account creation. Retweeted status has fields with original tweet metadata including its author profile data. Entities of a tweet contain some additional information about a message, such as lists of hashtags, urls, user mentions and symbols. These characteristics can also have an impact in the research. Retweet is a repost of another message of user of Twitter on someone's profile. With retweet information, it is possible to see how users interact and what information they share.

Data preprocessing: We perform various data cleaning steps before analysing the data. Firstly, we remove the special characters as they can impact features like length of the tweet as well the sentiment analysis. We also convert the time into Unix timestamp in seconds. For the missing values, observations that have most of their attributes empty are removed from the dataset. In other cases, if there is no information about one or few numeric features, zeros are inserted instead. For instance, user simply may not have any friends or followers. We also removed non-English tweets from out datasets. As expected, the collected datasets was very imbalanced, as there are a lot of messages that have zero retweets. In our dataset only 20% *Cryptocurrency*, 22% *Smartphone brands*, and 25% *Football* tweets were retweeted. In order to help prevent overfitting, we decided to downsample the majority class to make the same number of observations in each class, to avoid the problem of imbalanced dataset.

4 Methodology

In this section, we describe our proposed approach to information diffusion analysis and prediction. The number of retweets is the most indicative measure of information diffusion [5], thus, we use it as a target variable in our prediction models. We started with the regression approach of predicting the closest value for the retweets however, it is not easy to predict the exact number [5]. Thus, we performed two types of classification tasks. Firstly, the binary classification where we only predicted if a tweet will be retweet or not. Secondly, in multiclass classification, we created classes and predicted for a certain tweet to which types of a class it belongs. In addition, we normalized this data so that we predict the retweet number after the same period of time for each original tweet. In this case, period of time is the time range from the moment each message had appeared till some point of time in the future. The time period of 7 days was chosen for further analysis. Each tweet which has been retweeted several times has a different popularity window in time and eventually the retweeting process stops. Tweet lifespan of various tweets in our dataset shows scale-free pattern and the bigger set of tweets is concentrated below 50 hours duration which is around two days.

4.1 Class labelling

In order to apply any of the supervised machine learning models, depending on the type of classification, we need to label the data appropriately or in

other words, make classes from our numerical target variable. We created two classes: "retweeted" and "no retweet" for binary classification task, and 4 classes for multiclass task. In case of Binary classification, for each original tweet, we calculate the number of retweets for it. If the value is more than 1, we assign it "retweeted" otherwise "no retweet". For multiclass classification, we created 4 classes based on the number of retweets: Very Low (0 to 10 retweets), Low (11 to 90 retweets), Medium (91 to 170 retweets), High (more than 170 retweets). There are more samples with less retweets and few samples with high value of retweet count which are more difficult to predict.

4.2 Feature extraction

In this section, we describe various features that we considered for our prediction model.

1. User features: User profile information is very likely to be influential on how many times a tweet of a user will be retweeted. We selected the following most intuitive features:

- Followers count: Number of people who follow a user.
- Account age: Period of time calculated as difference between the time account was created and the time of tweeting a message.
- Listed count: Number of public lists that a user is a member of.
- Verified: Indicator if a user is verified or not. Binary variable.
- Friends count: Number of friends of a user.
- Statuses count: Number of tweets posted by a user.

2. Content features: There are many features that can be extracted from tweet's text. Some of them such as number of user mentions, hashtags and URLs lists are given in a tweet' metadata. Following is the list of all the content features:

- Tweet length: Number of symbols in a tweet, including spaces.
- User mentions: Number of user mentions with @ notation.
- Hashtags: Number of hashtags.
- URLs: Number of URLs.
- Exclamation and question marks: Number of exclamation and question marks.

3. Sentiment features Sentiment analysis identifies emotions, for example, a text could have positive or negative sentiment in it. To find the role of sentiments in tweet diffusion nine different emotions and sentiments were defined and were analysed for their presence in each tweet. Sentiment extraction was done using Syuzhet package. The list of the sentiment (and emotions) includes *negative, positive, trust, joy, anger, disgust, sadness, fear, anticipation, surprise.*

4. Initial behaviour features The initial behavior of a tweet could also be an influential parameter in predicting tweets's diffusion in the future. In other words, the hypothesis is based on the fact that if a tweet attracts a lot of retweet in some "initial time period" then it is highly likely to attract

more tweets in the near future before eventually the retweeting about the original tweet dies out. Following features are used for characterising the initial retweet behaviour:

- Current retweet count: Number of retweets of the message happened in the given initial time period
- Time alive of message: Period of time since the original message appeared till the last retweet in the given initial time period
- Tweet rate: Number of retweets in the time frame of one hour divided by time alive of message in the given initial time period
- Mean difference between retweets: Mean difference between retweets in the given initial time period
- Max difference between retweets: Max difference between retweets in the given initial time period

In addition, network of user followers can also contribute in tweet diffusion. Subfollowers are the number of people who follow users who retweeted an original tweet in the given initial time period. We added this feature using information of each user who have retweeted a message and added these features to an original tweet data.

4.3 Splitting and cross-validation

As for any standard prediction task in machine learning we split the data in training (80% of the data) and testing sets (20% of the data). Considering that the information propagation diffusion prediction task is oriented on future popularity of a message it seems more reasonable to split by time when tweets appeared. However, it was studied in [11] that there is no significant difference between random and chronological splitting methods for this particular prediction task. For our analysis we used k-fold cross-validation technique, which randomly splits the data into k-samples, and trains model multiple times so that each k-th fold serves as a test for the k_{th} instance. For our experiments we used k=10 for evaluating our model.

5 Evaluation

One of the important steps in working with predictive algorithms is to measure and compare obtained results. We evaluate our model using standard machine learning metrics such as confusion matrix, accuracy, F1-measure. We are not explaining these metrics due to space limitation and also, these metrics are well known in data science community.

5.1 Results

As described in the Section 4.1 instead of predicting the exact value of retweeets, we divided the total retweets into four classes. We used two bagging based methods namely 1) Random Forest and, 2)Treebag and two boosting based methods namely 1) Gbm (Gradient Boosting Machine), and, 2) Xg-

boost algorithms for the prediction tasks. We train the models using training data of cryptocurrency dataset and then evaluated it on the test set of cryptocurrency as well as on the other two datasts. The Table 2 shows the summary of results for multiclass prediction task.

Table 2: Performance results of classification

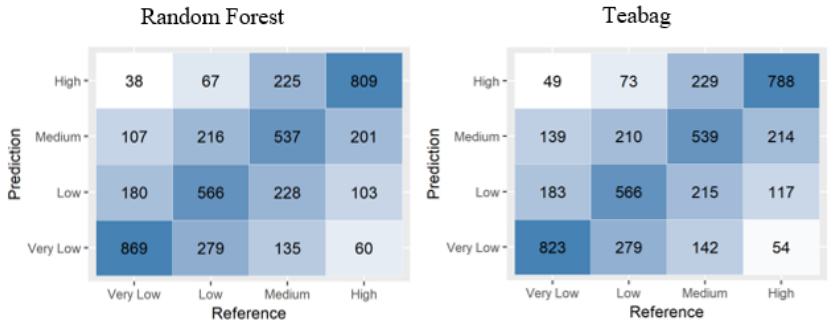
Feature set	Random forest		Gbm		Xgboost		Treebag	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
	Cryptocurrency dataset							
User + Content	0.6024	0.6741	0.4820	0.5830	0.5013	0.5991	0.5879	0.6816
User + Content + Sentiment	0.6037	0.6898	0.4822	0.5924	0.5110	0.6638	0.5907	0.6682
	Smartphone brands dataset							
User + Content + Sentiment	0.6011	0.6823	0.4696	0.581	0.5014	0.6532	0.5827	0.6777
	Football dataset							
User + Content + Sentiment	0.5799	0.6645	0.4683	0.5745	0.5010	0.6612	0.5789	0.6725

From the summary Table 2, we can conclude that Random Forest and Treebag algorithms performed better than Xgboost and Gbm. Random Forest showed the best on the performance metrics. Bagging algorithms confirmed their good performance by each class as well. Other algorithms showed worse performance overall (as seen from the Table 2) and by each class. However, the edge classes (High and Very low) are predicted quite good in all algorithms up to 72% of accuracy of one class. To look at the performance of each class we used confusion matrix (CM) which is a good visualisation for spotting the prediction errors. Figure 1 shows the CM for the two best algorithms.

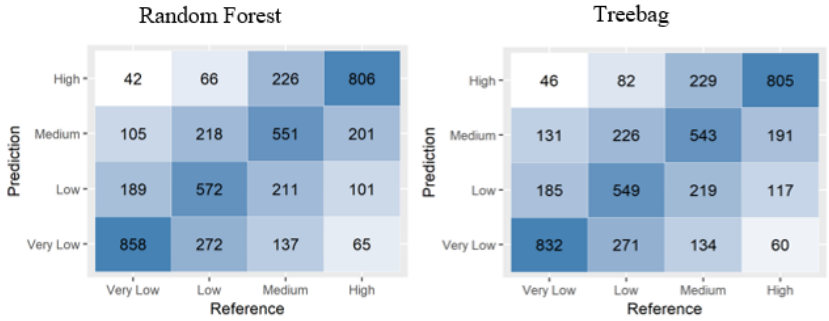
Feature importance: Since there are many performance metrics that can be obtained from confusion matrix (CM) we focus more on ones selected for this prediction task. Figure 2a shows F1-measure for each class and model and it is evident that Random Forest performs the best. However, looking at each class separately, we can conclude that they have not brought significant improvement. They could make a model more stable so we need to see the feature importance values. Figure 2b shows ordered importance of the best model of used features for the Random Forest. Other algorithms have their own order of important features that is not presented here due to space limitation but it is worth to note that first 5-7 features are the same across all the models (user related features) and group of first 8 features have relatively high level of significance comparing to others. These are 5 User features and 3 three Content features. In addition, positive sentiment is the most important from Sentiment feature set for the prediction.

5.2 Predicting tweet popularity using initial retweet behaviour features

This part of analysis requires usage of initial behaviour features introduced in the Section 4.2. These features most likely boost the performance for multiclass prediction as they could give more accurate result for each class of target

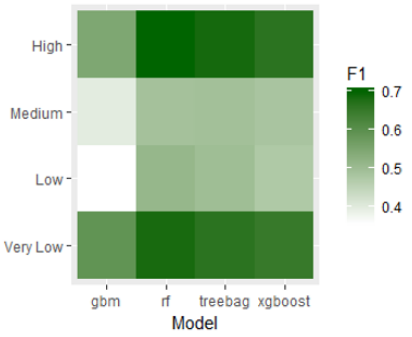


(a) User + Content features

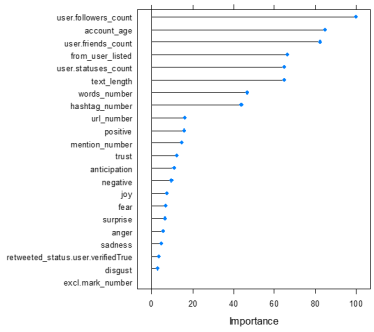


(b) User + Content + Sentiment features

Fig. 1: Confusion matrix



(a) Comparison of models F1-measure among all classes (User + Content + Sentiment features)



(b) Feature importance of Random Forest with User + Content + Sentiment features (top 20)

Fig. 2: Feature Analysis

variable. Moreover, it was studied how performance changes with increase in the initial time range. Analysing this, we obtained the point of time range after which there is no significant change in performance of prediction model.

We defined the following thresholds for the analysis: 1, 2, 3, 4, 5, 10, 30, 60 minutes (see Table 3). Since the Random Forest algorithm showed the best performance and this type of analysis is more complex than previous one, we decided to compare output for our datasets using only Random Forest.

Table 3: Performance results of multiclass prediction with initial behaviour features

Dataset	Cryptocurrency		Smartphone brands		Football	
Initial behaviour time range	Acc	F1	Acc	F1	Acc	F1
1 min	0.626	0.726	0.611	0.685	0.579	0.680
2 min	0.643	0.728	0.631	0.692	0.587	0.685
3 min	0.651	0.729	0.646	0.696	0.594	0.689
4 min	0.655	0.746	0.653	0.710	0.602	0.697
5 min	0.656	0.747	0.655	0.717	0.606	0.702
10 min	0.656	0.748	0.658	0.724	0.611	0.709
30 min	0.659	0.752	0.661	0.729	0.614	0.713
60 min	0.674	0.764	0.667	0.738	0.622	0.722

As expected, the performance increases with increasing the initial time range. Compared to 1 min, in 60 min period, accuracy increased by 4.8% and F1-measure by 3.8% in *Cryptocurrency* testing set. In addition, we can see that even 1 min initial behaviour features improves the accuracy of model by 2%. *Smartphone brands* and *football datasets* did not perform well but still we can observe similar improvement trend. Certainly, different datasets have different retweet activity. From the confusion matrices (Figure 3), we can see the improvement of prediction by each class. It is clearly seen that the initial time features strongly affect the result, especially detecting well Low and Medium classes that are more difficult to distinguish.

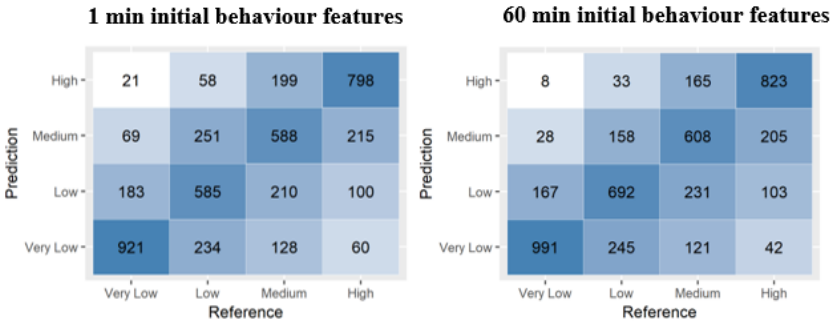


Fig. 3: Confusion matrixes of predictions using different range initial behaviour features in Cryptocurrency dataset

Feature importance: From the previous experiments we found out several most significant features for prediction. Considering the fact that initial time features improved the results we can see how the order of importance changes.

Obviously, current number of retweets gets the first place with increasing of the initial time range. Most of the content and user features like number of followers, number of friends, account age, if user listed or not remain important (in top 27 features). Some initial behaviour time features in addition to current count of retweets ranked higher positions (such as subfollowers) in 60 min time interval.

Another approach is predicting behaviour of certain message in the next period of time in the future. For this task we decided to use one-hour time frame. To evaluate it, we first organize the dataset so that the data is divided into one-hour time frames for each unique tweet. Target variable (retweet number) was taken using information of the next hour. In this task, the goal was to predict what would happen in the next hour based on the information about retweet activity from the previous hour. For this purpose, the same features are used as in previous approach of min by min analysis, but they were created from reorganized one-hour-frame data. Concerning Initial Behaviour features, they were created using whole one-hour time frame data for each original tweet. Therefore, they are used as retweet behaviour characteristics of previous hour. Training and testing sets were adjusted according to this task observations that already belong to class High in previous hour were removed. Therefore, we could see if observations of Very Low, Low or Medium class can move to class with bigger number of retweets. The Table 4 gives an overview of the results.

Table 4: Performance results of multiclass prediction of next hour based on previous hour

Dataset	Random Forest		Gbm		Xgboost		Treebag	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Cryptocurrency	0.571	0.678	0.513	0.567	0.522	0.581	0.543	0.662
Smartphone brands	0.563	0.669	0.504	0.558	0.519	0.580	0.538	0.650
Football	0.544	0.654	0.502	0.551	0.513	0.575	0.533	0.645

The performance of the models is worse comparing to multiclass models in previous sections. We can observe that it is difficult to predict tweet popularity using this approach even with four number of classes. This might be caused by the retweeting behaviour of our training data from *Cryptocurrency* dataset or due to not enough amount of training data.

6 Conclusion

In this work, we analysed Twitter messages and extracted various features for predicting the retweet trends. We extracted in total 27 features, broadly categorized into namely i) Content, ii) User, iii) Sentiment and, iv) Initial Behaviour and analysed their impact on model prediction results. We performed multiclass prediction task to understand approximately how much retweet a tweet can attract. Our approach showed decent performance using

Content, User and Sentiment features, that is an overall accuracy of 60%. We also analysed retweet behaviour in the first minutes of tweet existence and find out how it affects the prediction power of the model. We used all sets of features for this task and discovered that having information even of 5 minutes is enough to increase the overall accuracy value of 5%.

We have multiple future directions towards this work. We believe that larger dataset certainly would improve the stability and effectiveness of models. We would like to study if people retweet more when they see that a message is already popular. We would also like to include more features such as by including information about all friends and followers of a tweet's originator and by improving sentiment analysis by including emojis etc. We would also like to predict different measure of popularity such as predicting how many comments a tweet can get.

7 Acknowledgements

This work is supported by H2020 framework project, SoBigData.

References

1. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 u.s. election: Divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05, pp. 36–43. ACM, New York, NY, USA (2005)
2. Cazzoli, L., Sharma, R., Treccani, M., Lillo, F.: A large scale study to understand the relation between twitter and financial market. In: Network Intelligence Conference (ENIC), 2016 Third European, pp. 98–105 (2016)
3. Chen, J., Li, H., Wu, Z., Hossain, M.S.: Sentiment analysis of the correlation between regular tweets and retweets. In: 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), pp. 1–5 (2017)
4. Cohen, K., Johansson, F., Kaati, L., Mork, J.C.: Detecting linguistic markers for radical violence in social media. Terrorism and Political Violence **26**(1), 246–256 (2014)
5. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11, pp. 57–58. ACM, New York, NY, USA (2011)
6. Kafeza, E., Kanavos, A., Makris, C., Vikatos, P.: Predicting Information Diffusion Patterns in Twitter. In: 10th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), vol. AICT-436, pp. 79–89. Springer (2014). Part 3: Social Media and Mobile Applications of AI
7. Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., Kustarev, A.: Prediction of retweet cascade size over time. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, pp. 2335–2338. ACM, New York, NY, USA (2012)
8. Mazloom, M., Rietveld, R., Rudinac, S., Worring, M., van Dolen, W.: Multimodal popularity prediction of brand-related social media posts. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 197–201. ACM (2016)
9. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: Bad news travel fast: A content-based analysis of interestingness on twitter. In: Proceedings of the 3rd International Web Science Conference, WebSci '11, pp. 8:1–8:7. ACM, New York, NY, USA (2011)
10. Okubo, K., Oida, K.: A successful advertising strategy over twitter. Computer and Information Science **10**, 10–22 (2017)

11. Sarabchi, F.: Quantitative Prediction of Twitter Message Dissemination: A Machine Learning Approach. Master's thesis, Technical University of Delft (2015)
12. Shafiq, Z., Liu, A.: Cascade size prediction in online social networks. In: 2017 IFIP Networking Conference (IFIP Networking) and Workshops, pp. 1–9 (2017). DOI 10.23919/IFIPNetworking.2017.8264864
13. Taxidou, I., Fischer, P.M.: Online analysis of information diffusion in twitter. In: Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, pp. 1313–1318. ACM, New York, NY, USA (2014)
14. Wu, B., Shen, H.: Analyzing and predicting news popularity on twitter. Int. J. Inf. Manag. **35**(6), 702–711 (2015)
15. Yang, J., Counts, S.: Predicting the speed, scale, and range of information diffusion in twitter (2010)