IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Feedback Prediction for Blogs – How Machine Learning Can Stop Fake News Spread.

Harshit Bhatt[1],Gunjan Bahuguna[2]

Department of Computer Science, Birla Institute of Applied Science, Bhimtal

1. Abstract

Mankind have witnessed an unbelievable growth of the importance of social media in last few years.
When it all started tweets, facebook messages, youtube videos,
social tagging, blogs writing systems, etc. served more-less just as a source of entertainment of a few
enthusiastic users. Today the time has changed, news spreading over social media may be considered as the
most important changes of our society, Mostly the Bad news. Also advertisements and news about new
services, products related to the services and companies launching the products, these are spreading very quickly through the means of social media. This is indeed a great possibility for promoting
new products and services. But, according to sociological studies, negative opinions and news spread much quicker than positive news, therefore, if negative news appears on social platform about a particular organization, it must react quickly, in order to avoid losses.
Due to such large amount of documents appearing every second in social media platforms, analyzing
all these documents by human experts is nearly impossible, hence there is a
Colossal need for the *automatic* breakdown of such documents.
For the analysis of the documents, we have to consider some special properties of the application domain
Into account. In particular, the uncontrolled, dynamic and rapidly-changing
Content of social media documents

This study demonstrates how *Machine Learning* with *Data Mining* techniques can address the aforementioned challenges.
In this paper, we focus on the analytic components that allow to predict the number of feedbacks that a document is expected to receive in the next 24 hours.

2. Domain-specific concepts

In order to address the problem, first, we characterized some domain-specific concepts that are specified in this section. We say that a *source* produces *documents*. For example, on some blog writing site, new documents appear regularly, therefore, we say that the site is the source of these documents.
From the point of view of our work, the following parts of the documents are the most relevant ones:
(i) **Main text of the document :** The text that is written by the author of the document or blog etc., this text describes the topic of the document.
(ii) **Links to other documents:** Pointers to semantically related documents, in our case, trackbacks are regarded as such links.
(iii) **Feedbacks:** Opinions generated on the social media, generally referred as comments of the users about a particular text is very often expressed in form of feedbacks that the text receives.
Feedbacks are textual comments expressing reactions and to the original text of the document posted.
All these aforementioned entities are relevant for the task.
Therefore, we excerpt *time-stamps* for these entities and store all the relevant data together with these timestamps.

3. Feedback prediction

Feedback prediction is scientifically the most interesting component of the Study, therefore we focus on feedback prediction.

3.1. Problem Formulation

Given some blog documents that arose in the past, for which we already Know when and how many feedbacks they have received, the task is to predict how many feedbacks *recently* published blog-entries will receive in the next, say *'H'* hours. We consider the blog documents published in the last 72 hours as recently published ones, and accordingly, we set H = 24 hours.

3.2. Machine Learning for Feedback Prediction

We use *machine learning* to solve this prediction problem, in particular *regression models* are used. Here we consider the instances as the recently posted blog documents and the target is the number of feedbacks that the published blog will receive in the next '*H'* hours. '*H*' is the variable amount of time we consider for analysis.

Most regression algorithms assume that the instances are vectors. Additionally, it is assumed that the value of the target is known for some instances, and based on this information, we want to predict the value of the target for those cases where it is unknown.

First, using the cases where the target is known, a prediction model(*regressor)*, is produced. Then, that regressor is used to predict the value of the target for the instances in which the target is unknown.

Now, in order to apply machine learning to the above discussed prediction problem, we have to consider the following issues:

(i) We need to convert the instances (blog documents) into vectors, and

(ii) We need some previous data from some model for which the value of the target is already known (**Train data**).

Considering the first problem, i.e., for turning the documents into vectors, we extract the following features from each document:

1. Fundamental features: Number of feedbacks in the previous 24 hours In relation to the baseTime; number of feedbacks in the time interval from 48 hours prior to baseTime to 24 hours prior to baseTime; how the number of feedbacks increased or decreased in the past time in relation to the base time; number of feedbacks in the first 24 hours after the publication of the blog, but before the baseTime;

2. Textual features: The most discriminative lot of words features. The good words, the bad words etc.

3. Weekday features: The features that describe that in a particular week, on a particular day, the main text of the document was published and for which day the prediction has to be determined.

4. Parent features: We consider a document *dP* as a parent of document *d*, if *d* is what we get as a reply to *dP* , i.e., there is a trackback link on *dP* that points to *d*; parent features are the number of minimum, maximum and average number of feedbacks that the parents received.

We solve the first problem as follows: We select a particular date and time in the past and pretend as if the current date and time would be the selected date and time. We call the selected date and time *baseTime*. As we actually know what happened after the baseTime, i.e., we know how many feedbacks the blog entries received in the next '*H'* hours after the *baseTime*, we know the values of the target for these cases. While doing so, we only take blog pages into consideration that were published in the last three days relative to the baseTime, because older blog pages usually do not receive any more new feedbacks.

A similar approach allows us to significantly evaluate the prediction models: we choose a time interval, in which we select different times as base-Time, calculate the value of the target and use the resulting data to train the regressor. Then, we select a disjoint time interval in which we again take several baseTimes and calculate the true values of the target. However, the true values of the target remain hidden for the prediction model, we use the prediction model to estimate the values of the targets for the second time interval. Then we can compare the true and the predicted values of the target.

5. Conclusion

In the last decade, the importance of social media grew unbelievably. Here, we presented a proof-of-concept social media analysis. In particular, we aimed to predict the number of feedbacks that blog documents receive. Our study will allow us to drag data from blogs and perform predictions. On the other hand, the study shows that there is still room for improvement, while developing new models for the blog feedback prediction problem seems to be a non-trivial task: with widely-used techniques, in particular ensemble methods, we only achieved marginal improvement. In order to motivate research in this area of growing interest, we must create some software prototypes that would predict the feedbacks using train-data.

6. References

- https://archive.ics.uci.edu/ml/index.php
- https://archive.ics.uci.edu/ml/datasets/BlogFeedback
- http://www.businessdictionary.com/definition/feedback.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4685608/
- https://www.quora.com/What-is-a-training-data-set-test-data-set-in-machine-learning-What-are-the-rules-for-selecting-them