

A Supervised Method to Predict the Popularity of News Articles

Ali Balali, Masoud Asadpour, Hesham Faili

University of Tehran, School of ECE,
College of Engineering, Tehran,
Iran

{balali.a67, asadpour, hfaili}@ut.ac.ir

Abstract. In this study, we identify the features of an article that encourage people to leave a comment for it. The volume of the received comments for a news article shows its importance. It also indirectly indicates the amount of influence a news article has on the public. Leaving comment on a news article indicates not only the visitor has read the article but also the article has been important to him/her. We propose a machine learning approach to predict the volume of comments using the information that is extracted about the users' activities on the web pages of news agencies. In order to evaluate the proposed method, several experiments were performed. The results reveal salient improvement in comparison with the baseline methods.

Keywords. Text mining, comments volume, content popularity, user behavior, social media.

1 Introduction

In recent years, the number of users that use social media on a daily basis is dramatically increasing according to Pew Research Center¹, 65% of U.S. adults use them in 2015. Consequently, the number of comments published about these media has increased a lot. These comments express users' feelings, opinions and ideas about their individual or social issues, commercial products, etc. These comments can be studied by social organizations, governments, or commercial bodies, in order to extract valuable information for social welfare management, planning and marketing.

Among the comments, those that are left for news articles are of great importance, since they express the feeling and opinion of people about current social issues. We would like to identify

which news articles influence the public. Which features among all features of articles are important for people so that they are encouraged to read them and above that encourage them to express their opinion by writing a comment.

The volume of comments (VOC) received by a news article shows its importance to public. It also indicates its diffusion rate and impact on the society. Controversial articles show important issues for public. When a visitor leaves a comment on a news article, this indicates not only he/she has read the news article but also the news has been important to him/her.

Prediction of VOC can be useful in specifying a suitable place for that article on the website or specifying suitable advertisements along with proper advertisement price. It can also be used as a part of Business Intelligence (BI) software for news agencies. It can help editors to change the article for maximum attraction. Also, it can propose an appropriate order for publish. The proposed method can also be useful in sorting and filtering comments based on their importance.

The main goal of this paper is to propose an approach to predict VOC based on the textual and temporal features of news. Very few studies have been carried out on this topic. This problem can be modeled as a classification (to predict VOC) or regression classification (to predict the number of comments). The datasets that we use in this paper were collected from several online news agencies of Iran. They daily publish news articles about current events in Iran and abroad in different categories such as politics, economy, culture, society, technology, sport, etc.

¹ <http://www.pewresearch.org/>

In this paper, we also define several features for weighting words based on users' interests that can be useful in devising slogans for electoral campaigns and advertisements. Also, it can be used as a word weighting method in various areas such as Information Retrieval.

The main contributions of this work are as follows.

- We propose a supervised approach to predict the volume of received comments on news articles prior to its publication. This approach shows people's interest to social issues.
- The proposed method focuses solely on the lexical level so it can be used on all online news agencies regardless of their languages.
- To extract various factors that attract users, we define four novel textual and non-textual feature sets. These features are extracted from the users' activities on web pages. The textual features can weight words based on users' interest over time.
- The evaluation results reveal higher accuracy in comparison with the baseline methods.

The rest of this paper is organized as follows. In Section 2, we describe related works. Then in the next section, we explain the proposed method and introduce the features we use. In Section 4, the experiments, datasets, evaluation metrics, and experimental results are presented. Finally, conclusion and future works come in Section 5.

2 Related Works

Various internal and external factors affect the number of comments that news articles might receive. There can be external factors such as community [1] or reputation [2] of the agency or its web site that effects the number of comments. These factors are out of the context of this paper. In this paper we focus only on the internal factors like content and temporal information.

Prediction of VOC can be done both before publication [4, 5, 23, 24] and shortly after publication [3]. The latter can be more accurate as it could receive some actual information like the current number of visitors, the number of comments received on the published article so far and so on. In this paper we focus on the former

case that is naturally more difficult due to missing some actual information.

In the following section, the researchers done on both cases are described.

2.1 Prediction of VOC Before Publication




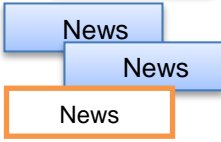
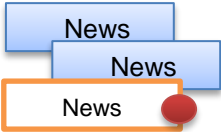

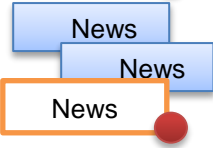
Tsagkias et al. [4] introduced several feature sets which affect VOC: Surface features like month, day, and hour of day; Textual features such as repetition frequency of words, and length of posts; Semantic features like Named Entity Recognition using name of people, organizations and places; and finally real-world features such as weather conditions. To combine these features, they used two binary classifiers. The first classifier predicts whether a news article receive comment or not. The second classifier predicts VOC (low or high). The dataset is prepared from several news agencies of Netherland. This method is considered as baseline in this paper.

Ren et al. [23] used machine learning techniques to predict the popularity of online news. The popularity measure was the number of shares under a news article. They extracted 59 features describing different aspects of each article. They implemented 10 different machine learning models that Random Forest got the best result.

Very few studies directly address the problem of prediction of VOC. In the following, we explain some papers that focus on ranking and filtering news comments that could be useful in prediction of VOC.

Siersdorfer et al. [5] analyzed YouTube comments and predicted comment ratings. They showed that the sentiment and the category of comments could affect the rating of comments. According to their research, comments related to music category usually have higher rate than comments in other categories. Also, they showed that comments, which use positive words, usually, receive better ratings. The positive or negative score of a word is obtained by SentiWordNet. Finally, to predict the rate of comments, SVM classifier was used. The used features are terms with their weights. One of the applications of this method is for automatic comment approval i.e. confirmation or rejection of comments. Some other papers [6, 7] have done similar works in this area.

Table 1. The process of the proposed method (For each news, the steps 2 to 6 is repeated)

| Step | Test data | Event | Train data | Description |
|------|-----------------------------------------------------------------------------------|----------------------------------------------|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| 1 | | 2 months for learning the model |  | The method is learning the model. |
| 2 |  | The news articles are published |  | The method predicts the volume of comments. |
| 3 | | |  | The news article is added to train data. The textual information is extracted. Features are updated. |
| 4 | | Comments are published |  | The textual information is extracted from the comment. Features are updated. |
| 5 | | After 24 hours of published the news article |  | The number of visitors are extracted. Features are updated. |
| 6 | | After 24 hours of published the comment |  | The number of like, dislike and replies are being extracted. Features are updated. |

Hsu et al. [6] proposed a method to rank news comments. The dataset is prepared from Digg website. According to this method, the high quality comments are placed in the top of the list of comments. The proposed features for comment ranking is useful for our case.

They proposed several feature sets such as comment visibility, user reputation and some content-based features. They used SVM classifier to combine these features. They mention that the scores are biased with publication time. Comments left earlier for a news article, are likely to get more scores.

They solved the bias problem using “boosted ranking method”. Some other papers have also focused on removing the biased score of comments [8].

2.2 Prediction of VOC after Publication

Jamali et al. [9] proposed a method to predict the rate of news articles. The dataset is prepared from Digg website. Users in Digg can rate news articles similar to rating comments. The purpose of this paper is predicting the news rate based on the rate of comments, the number of comments, the reply-tree structure and the news category. They have used several classifiers like decision tree, SVM, KNN and SVM regression.

Tatar et al. [10] and Tsagkias et al. [11] proposed a method which predicts the number of new comments after publication of an article. This method predicts the final number of received comments after observing a news article for a short period. Tatar et al. method used linear regression to predict the number of comments. Their dataset

was prepared from two news agencies in France. As observation period is relatively long, precision of prediction is good. Tsagkias et al used a dataset collected from 8 news agencies of Netherland.

2.3 Information Extraction from Comments

In this section, we emphasize on the importance of comments. Since, comments are usually short, informal, and complex to be understood, few works have been done on them [12]. Rizos et al. [22] presented a framework for the prediction of news story popularity. The framework focusing on features from two sources of social interactions inherent in online discussions: the comment tree and the user graph. Even, the reply tree of comments includes some valuable information such as showing important comments [13]. Also, the user's comments are used in different applications like user behavior analysis, finding important users [14] and post summarization [15].

3 Our Proposed Method

Our method is a machine learning approach for prediction of VOC. We need a set of features to be used in learning. We extract four textual and non-textual feature sets from the train data. The textual features are extracted from the body of news articles and comments. Table 1 summarizes the steps we follow. First, the method is learned on two months of data. After this time, as soon as a news article is published, VOC is predicted for it. Then, its textual information is imported to the train data.

Some features such as number of likes and dislikes, structure of the reply-tree of comments and the number of predicted visitors are imported to the train data after 24 hours. The 24 hours threshold is to provide enough time for monitoring user activities. The threshold can be decreased by continuous fetching of the web pages.

The main challenge is finding good features that are correlated to the attention of visitors. In the next subsection, the classifiers and predictors are explained. Then, the features set are proposed.

3.1 The Classifiers and Predictors

We assume three types of models and try to learn the models in order to see which one leads to best results.

- First, we model the problem as a binary classification task: "with comment" and "without comment". The "without comment" class shows diffusion rate of the news article is low, so it does not affect the society that much. The second class "with comment" shows news article has a good diffusion and it spreads in the society.
- In the second model, we classify the number of comments into three classes "No comment", "Moderate" and "High". We use Random Forest Classifiers to learn both models.
- In the third model, we assume it as a regression classification task. The purpose of this model is to predict the exact number of comments before publication. We use Linear Regression Classifiers and Ensemble Learning to predict the number of the comments for each news article.

Random Forest is an Ensemble Learning method that is useful for multiclass problems with balanced data distribution. This classifier constructs several decision trees and specifies the mode of the classes for aggregation of the results [16].

Linear Regression Classifier is a simple classifier that has a perceptible linear model. This classifier extracts the relation between a set of independent variables with one dependent variable. The best model is created based on a line optimizing residual sum of squared errors metric [17]. This metric tries to minimize the distance between the set of samples in the train data and predicted line. The machine learning-based approach that is used in this paper is available in Weka software [18]. The used features are introduced in the next section.

3.2 Features

We have used four feature sets. Two feature sets focus on textual information. Textual information of news articles and comments are based on unigram and bigram language models. The unigram and

bigram terms that have low frequencies are usually typo errors. In our case, the unigrams that appear less than 100 times and the bigrams that appear less than 10 times are removed from the list. In the following, the four feature sets are explained in detail.

3.2.1. Global Textual Features

Some terms (unigram or bigram words) are always important for people such as celebrities or sport teams. To extract these terms, we define a set of global textual features identifying them. The train data is updated over time. Each feature rates news articles based on its constituent words. Each term is given a value according to its importance.

Each textual feature calculates two values for an article based on its title and body. Title and Content-scores are calculated according to (1) and (2). To compute Content-score, we select only 10 distinct terms from the body of news article that have maximum value:

$$Title - score = \frac{\sum_{w \in Title} Score(w)}{|Title|} \quad (1)$$

$$Content - score = \sum_{\substack{w \in \text{body} \\ (score(w) \in \text{distinct}(\text{content})), Desc}}^{10} score(w) \quad (2)$$

Here $score(w)$ is the importance of word w . Global textual features are introduced below.

- The number of replies.

The number of replies for a comment shows its importance. Some websites show replied comments as a tree structure and some as a list sorted in chronological (or reverse) order. Some papers work on automatic reconstruction of the tree structure from a list [13, 19, 20].

This feature calculates the value of each term based on the average number of received replies for the comment that includes it.

- Number of likes/dislikes.

In some websites, users can announce their agreement or disagreement with a comment by clicking on the like / dislike buttons of comments. The number of likes / dislikes is biased to its

publish time [6]. Comments that are left earlier usually have more likes / dislikes even if comments have the same bodies. We have evaluated several approaches to remove this bias. The best result was obtained by generalizing the number of likes / dislikes into some intervals. The value of each term is calculated according to (3):

$$Score(W) = \frac{\sum_{W \in \text{comments}} \begin{cases} \text{If } Like_c \leq \frac{Avg_L}{2} \rightarrow +0 \\ \text{If } Like_c > \frac{Avg_L}{2} \text{ and } Like_c \leq Avg_L \rightarrow +1 \\ \text{If } Like_c > Avg_L \text{ and } Like_c \leq 5Avg_L \rightarrow +2 \\ \text{If } Like_c > 5Avg_L \rightarrow +3 \end{cases}}{DF_c(W)} \quad (3)$$

Where DF_c is the frequency of the term W in all comments in the training data and $Like_c$ is the number of likes. Also, Avg_L is the average number of likes in the whole training data. The dislike score is calculated similar to like.

- The number of received comments.

Based on this feature, leaving a comment for a news article increases the value of its terms. The publish time of comments are assumed to be available. So, we can consider comments as a stream sorted chronologically.

According to (4), the value of each term is the average number of received comments on the articles that contain the term in their body. If a term appears in a news article which could not receive any comments, its value is decreased. Two variables are defined for each term. The first one refers to the number of comments that contain a term and the second variable refers to the number of news articles that contain it.

In order to decrease the effect of outliers (e.g. articles that receive huge number of comments), according to (5), we can calculate the number of times a news article that include a term, receives comments more than average. This value is normalized with the term frequency of the term:

$$Score(W) = \frac{\sum_{d \in N(w)} C(d)}{|N(w)|} \quad (4)$$

$$Score(W) = \frac{\sum_{d \in News} \begin{cases} 1 & \text{If } C(d) > AVG_c \\ 0 & \text{otherwise} \end{cases}}{|N(w)|} \quad (5)$$

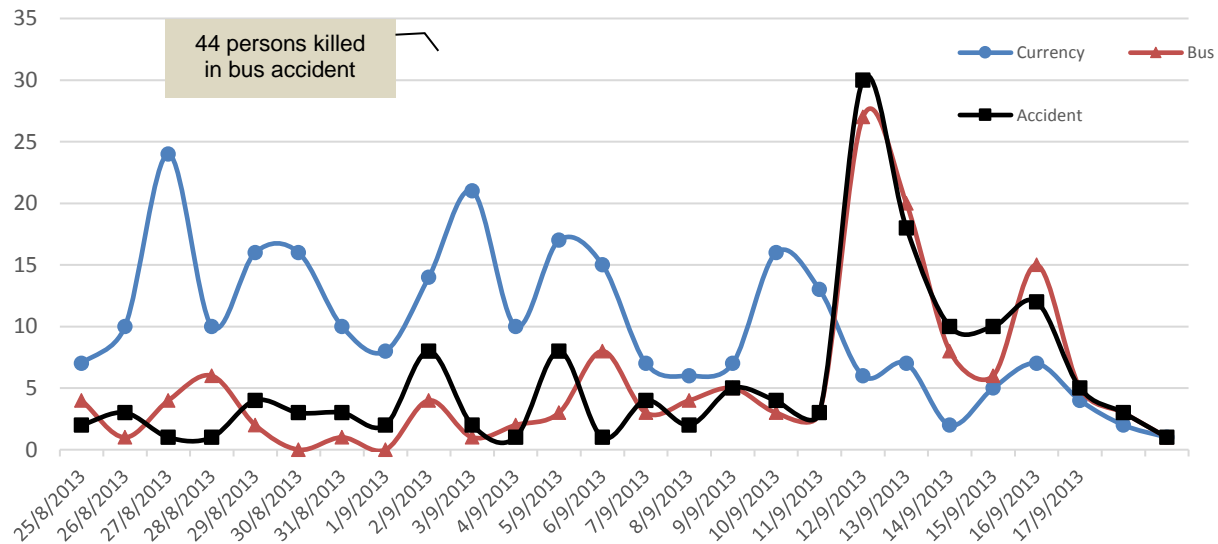


Fig. 1. The frequency of three words “Currency”, “Bus” and “Accident”

Where $C(d)$ is the number of comments received by news article d , and $N(w)$ is the number of articles that include term w .

- Page visit count.

Page visit count shows the number of users that have visited the article. The value of each term is calculated similar to the number of received comments i.e. (4).

- Discovering the news nature.

This set of features extracts some characteristics of the news based on its title. In the following, three binary features are defined that show different characteristics of an article:

- Interrogative nature: If question mark appears in the title, the value of this feature is set.
- Surprise nature: If exclamation mark appears in the title, the value of this feature is set.
- Polling nature: Some news articles try to collect visitor’s opinion. They encourage users to leave a comment and express their opinion about an issue. These articles usually receive large VOC. If the term “leave a comment” or “Your Opinion” appear in the title, the value of this feature is set.

3.2.2. Temporal-Textual Features

Terms might be hot in some period e.g. terms related to elections. In this feature set the term frequency on the articles of a day and the day before is counted. Stopwords are removed first.

Figure 1, shows the frequency of three terms “currency”, “bus” and “accident” over a short period of time. The “currency” term is a hot term in different days, whereas two “bus” and “accident” are not hot until 2013/9/11 when 44 persons were killed in a bus accident. The values of these two terms are increased suddenly after this date. This sudden increase shows an important and controversial event in real world. In the following, features in this category are described.

- Scoring terms based on temporal-textual information of comments.

The stream of comments in the training data is assumed to be sorted chronologically. We considered the term frequency of comments both for the current day and the day before. Some terms have always high frequency e.g. “President”. But the purpose of this feature is not these terms. The main goal of this feature is extracting the terms that suddenly become hot. So, the average frequency

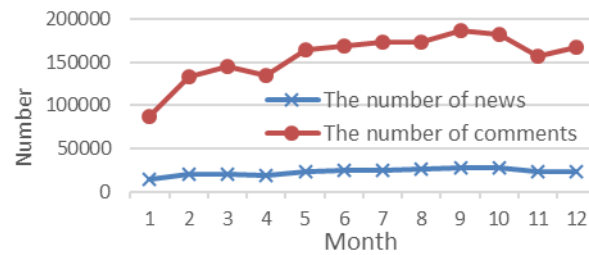


Fig. 2. The number of news and comments from 2010 to 2013

of each term is calculated on different days. If the term frequency is more than its average frequency on a special day, a reward is given to it. The reward is equal to difference between the term frequency in that special day and the average term frequency. For example, the term “accident” has the average frequency of 5, but in the mentioned date, its term frequency reaches to 30 (25 units above average). So, 25 point is added to its term frequency, summing up to 55. If the term frequency is lower than the average term frequency in a day, no reward is given.

- Scoring terms based on temporal-textual information of news articles.

Similar features could be extracted from the title and body of news articles. These features in fact reveal the importance of the terms according to news agencies and their editors.

3.2.3. Temporal Features

Temporal features are important in prediction of VOC. In our training data, articles published on 10 a.m. receive usually more comments than the other hours. Figure 2, shows the number of news articles and comments received within different months of the year. The minimum occurs at the first month of the Persian calendar (March 21-April 21, start of spring when the new year begins in Persian countries) when people go for holidays.

This set includes the number of comments received at a specific hour, day, month, year and the day of the week. To calculate the features for example for “hours”, we calculate the number of comments received in that specific hour divided by the total number of comments received. We calculate another features by dividing the number of comments received in that specific hour divided

by the number of articles published during that time. The other features related to days, months, years and the days of week are calculated similarly.

The features we have defined so far are not specific to any news agency. However, it is evident that the VOC is directly proportional to reputation of news agencies in general. We need some features to reflect the reputation of agencies. The best choice is the number of visitors of an article, but this information is not available before publication. Instead, we use two features that are extracted one day and one week before publication:

- The number of comments received on the previous day.

Sum of the number of received comments in the day before is used as a feature. This information is available in the train data. This feature might be insufficient in some occasions e.g. if the previous day is a special day like holiday. So, the second feature helps in this issue.

- The number of comments received on the previous week.

This feature calculates the sum of the number of received comments on the previous week. This feature is more robust to fluctuations in the number of visitors due to e.g. holidays.

3.2.4. Surface Features

This feature set focuses on extracting the surface information from news articles such as the number of terms in the title and body, its category and, existence of featuring image or movie for the article. If there is a term like “photo”, “image”,

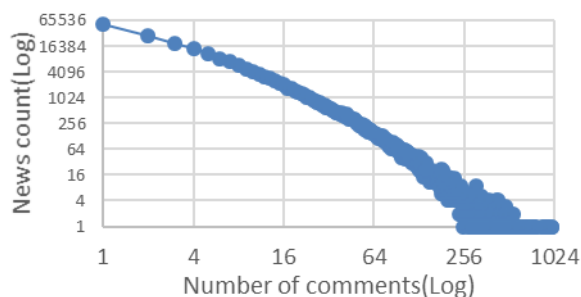


Fig. 3. The distribution of the number of comments

“movie”, on the title, it means the article is mainly focused on the attached photo or movie.

4 Experiments

In this section, we provide some details about the datasets we used for evaluation purpose and describe the evaluation metrics. Then, we present the results of the proposed method and compare them with the results of the baseline approach.

4.1 Dataset

We have performed our experiments on four popular online news agencies in Iran: Tabnak², Khabaronline³, Mashreghnews⁴ and Seratnews⁵. These are four online news agencies that have the highest traffic rank among online news agencies in Iran according to Alexa⁶. Besides popularity, the wide range of categories that are covered by them and their multilevel commenting structure are the other reasons for choosing them.

We crawled the websites completely. Then, we parsed the html pages and normalized the text by unification of ASCII codes, stemming and, removing html tags from the text. During this process we faced some challenges that are worth mentioning:

The editors and admins of online news agencies are free to select which article can receive comment and which articles cannot. In the collected dataset, we found many articles that have

large number of visits but have no comments. They are the articles for them commenting had been disabled. These articles create trouble for the learning process, as receiving no comment does not mean they are not important for the visitors. To solve this problem, articles that have a large number of visitors (higher than a threshold) but have not received any comment, are removed from the dataset.

The visit threshold is set based on the articles that have only one comment. To select appropriate threshold, news with one comment are selected and then, they are sorted ascending according to the number of visits. Then the first quartile is selected as the threshold. The threshold was set to 2100.

Some news agencies reject some comments that are in contrast with their viewpoints. So, we chose online news agencies that publish diverse political views.

Dataset has the following statistical features. We totally crawled 324,925 and 2,377,861 comments out of which 114,144 (35%) have no comments. The distribution of the number of comments is shown in Figure 3. According to this figure, the number of comments has power law distribution.

4.2 Evaluation Metrics

To evaluate the results, we use Correct Classification Rate (CCR) metric. This metrics

² www.Tabnak.ir

³ www.Khabaronline.ir

⁴ www.Mashreghnews.ir

⁵ www.Seratnews.ir

⁶ Alexa– The web information company, www.alexa.com

Table 2. The experiment results of the proposed method for binary classes

| Classes | Precision | Recall | F-measure |
|------------------|-----------|--------|-----------|
| Without comments | 0.87 | 0.873 | 0.871 |
| With comments | 0.93 | 0.928 | 0.929 |
| Weighted Avg | 0.909 | 0.909 | 0.909 |

Table 3. The experiment results of the baseline for binary classes

| Classes | Precision | Recall | F-measure |
|------------------|-----------|--------|-----------|
| Without comments | 0.638 | 0.65 | 0.644 |
| With comments | 0.833 | 0.826 | 0.829 |
| Weighted Avg | 0.77 | 0.769 | 0.77 |

Table 4. The experiment results of the proposed method for three classes

| Classes | Precision | Recall | F-measure |
|-----------------------------|-----------|--------|-----------|
| No comment | 0.876 | 0.921 | 0.897 |
| Moderate number of comments | 0.679 | 0.781 | 0.727 |
| High number of comments | 0.536 | 0.275 | 0.363 |
| Weighted Avg. | 0.746 | 0.762 | 0.746 |

Table 5. The experiment results of the baseline approach for three classes

| Classes | Precision | Recall | F-measure |
|-----------------------------|-----------|--------|-----------|
| No comment | 0.611 | 0.694 | 0.65 |
| Moderate number of comments | 0.592 | 0.677 | 0.632 |
| High number of comments | 0.389 | 0.19 | 0.255 |
| Weighted Avg. | 0.554 | 0.577 | 0.556 |

shows the proportion of correctly predicted classes. Also, to evaluate regression classifiers, we use Pearson's product-moment correlation coefficient and Root Mean Squared Error (RMSE). The Pearson Correlation Coefficient (PCC) shows the linear correlation of two quantitative variables. Also, RMSE shows the average difference between the predicted values and the gold data.

4.3 Experimental Results and Analysis

We implemented Tsagkias and et al.'s approach [4] as the baseline. It is the closest approach to our

work. To have a fair comparison, our method and the baseline approach have used the same classifiers in all experiments. To minimize bias, 5-fold cross-validation has been used in all experiments. Next, we show the results for binary classes.

4.3.1. Binary Classes

We divided the news articles into 2 classes: "Without comments" and "With comments". We used Random Forest Classifier to predict the classes of news articles. The precision, recall and F1-score for both classes of our method and the

Table 6. The bins confusion matrix by Linear Regression Classifier

| Predicted | | | | | | | |
|-----------|-------|--------|-------|--------|-------|-------|--|
| Gold data | First | Second | Third | Fourth | Fifth | Sixth | |
| First | 0.48 | 0.431 | 0.077 | 0.011 | 0.001 | 0 | |
| Second | 0.069 | 0.47 | 0.355 | 0.098 | 0.007 | 0.001 | |
| Third | 0.017 | 0.323 | 0.453 | 0.191 | 0.015 | 0.002 | |
| Fourth | 0.012 | 0.237 | 0.448 | 0.273 | 0.027 | 0.003 | |
| Fifth | 0.009 | 0.173 | 0.421 | 0.344 | 0.044 | 0.01 | |
| Sixth | 0.005 | 0.116 | 0.343 | 0.401 | 0.089 | 0.048 | |

Table 7. The bins confusion matrix by Ensemble Regression Classifier

| Predicted | | | | | | | |
|-----------|-------|--------|-------|--------|-------|-------|--|
| Gold data | First | Second | Third | Fourth | Fifth | Sixth | |
| First | 0.87 | 0.086 | 0.035 | 0.007 | 0 | 0 | |
| Second | 0.085 | 0.454 | 0.339 | 0.111 | 0.009 | 0.001 | |
| Third | 0.042 | 0.293 | 0.424 | 0.218 | 0.019 | 0.002 | |
| Fourth | 0.039 | 0.21 | 0.413 | 0.298 | 0.034 | 0.004 | |
| Fifth | 0.036 | 0.149 | 0.379 | 0.365 | 0.059 | 0.01 | |
| Sixth | 0.041 | 0.093 | 0.296 | 0.413 | 0.104 | 0.051 | |

Table 8. A summary of the achieved results by different classifiers

| Evaluation Metric Method | RMSE | Pearson Correlation Coefficient | CCR |
|------------------------------|------|---------------------------------|------|
| Ensemble classifier | 0.98 | 0.61 | 0.58 |
| Linear regression | 1.04 | 0.54 | 0.43 |
| MLP | 1.28 | 0.42 | 0.37 |
| Baseline (Linear regression) | 1.15 | 0.28 | 0.3 |

baseline are shown in Tables 2 and 3, respectively. The results show we can predict whether an article will receive comment or not with about 91% precision while the baseline has around 77% accuracy. Also the recall and F-measure of our method is higher. The CCR of our proposed method is equal to 90.86%, while the CCR of the baseline is equal to 76.90%.

4.3.2. Three Classes

We divide the articles based on their VOC into three classes “No comment”, “Moderate” and “High”. The articles that had not received any comments were put in “No comment” class. Then,

the average number of comments for those articles that had received some comments was set as the threshold for dividing the articles into “Moderate” and “High” classes. Again, we use Random Forest Classifier like the binary case. The results of our method and the baseline are presented in Tables 4 and 5 respectively. CCR of our method is 76.22% while CCR of the baseline is 57.74%.

It is seen that our method has around 20% better performance than the baseline approach in all three measures. It is also seen that, as the number of classes is increased, the difference between the results of the proposed model and the baseline increases.

4.3.3. Prediction of the Number of Comments

In order to predict the number of comments received for each article we use Linear Regression and Multilayer Perceptron (MLP) classifiers. As moderators reject some comments, even gold data is some time inaccurate. Thus, we predict a range for the number of comments. We divide the gold data into several bins based on the number of comments. The number of bins and their lengths depend on user requirements. At first, the number of comments is predicted for each article. Then, the predicted number is used to determine the bin.

We consider 6 bins; therefore, the predictor assigns a score between 1 to 6 to each article. The first bin includes the articles that have not received any comment. The 2nd bin includes articles that have received from one to the average number of comments (1x). For the remained bins except the last one the average is multiplied by two each time i.e. the 3rd interval includes articles that have received between the average (1x) and two times of the average (2x). The 4th interval includes articles that have received from 2x the average to 4x the average, etc. Finally, the 6th interval includes articles that received between 8x the average to maximum number of comments. This way of defining bins is good for balancing the number of articles in all bins.

By applying Linear Regression Classifier, we got PCC=0.54 and RMSE=1.04. Using MultiLayer Perceptron (MLP) Classifiers we got PCC=0.42 and RMSE=1.28. This means, there is a linear relation between the value of the features and the importance of the articles. The Linear Regression leads to better performance than MLP.

Using the baseline method, we got PCC=0.276 and RMSE=1.153. There are few textual features in the baseline that makes it perform weaker than our model. The strength of our method is choosing the textual features that have information on influence of the article.

The confusion matrix for bins is shown in Table 6. Fortunately, incorrect predictions usually end up in neighbor bins (so the predictions are not completely wrong). For example, the articles that belong to the 2nd bin are usually predicted to be in the 2nd or 3rd bin.

According to Table 6, the predictor predicted the first bin ("No Comment") with 48% accuracy.

However, it predicted the 2nd interval wrongly, instead of the first interval, in 43% of the times. According to Table 2, the binary predictor could detect these two classes with 90% accuracy. This made us to think of a combined solution; we used an Ensemble Classifier to combine Random Forest and Linear Regression Classifier.

We used Random Forest Classifier to predict binary values; whether a news article receives any comment or not. If the Random Forest predicts the article belongs to "No Comment" class, it is put in the first bin; otherwise, the exact bin (2nd to 6th) is specified by the Linear Regression Classifier.

PCC of the Ensemble classifier is 0.61 and its RMSE is 0.98. The performance is quite better in comparison with the other classifiers. Also, the confusion matrix of the Ensemble Classifier for different bins is shown in Table 7. According to this table, CCR is equal to 0.58. The ensemble predictor predicted the first interval ("No Comment") with 87% accuracy; this makes it better than the Linear Regression Classifier.

In Table 7, the darker cells show correct predictions and the bold cells show cells bigger than the correct cells. Fortunately, these incorrect predictions occur near the correct cells. The ensemble predictor can predict the 1st, 2nd and 3rd intervals well, but it faces trouble in prediction of the 4th to 6th intervals. There are several reasons for this: The number of instances in these bins is lower than the other three bins; this makes the classifier to perform better on the first 3 bins. If, also, we take a look at the tail of the histogram of the number of comments, we see that the tail (where the last 3 bins are located) is noisy. This noise is problematic in prediction.

The tail of the histogram belongs to the articles that have high VOC. They are mainly breaking news such as accidents, terrorist attacks, and sport events. Accurate prediction of them needs extra information such as presentation quality of the article (e.g. the location of the article on the webpage) or other information that are not easy to acquire.

Table 8 shows a summary of the results achieved by different classifiers. As it is seen, the Ensemble Classifier has better result than the other classifiers based on all measures.

Table 9. Evaluation of the features based on 3 classes model

| Metric Feature set | Precision | Recall | F-measure | Accuracy |
|--------------------|-----------|--------|-----------|----------|
| Temporal-textual | 0.52 | 0.552 | 0.523 | 0.552 |
| Global-textual | 0.627 | 0.65 | 0.631 | 0.65 |
| Surface | 0.663 | 0.681 | 0.67 | 0.681 |
| Temporal | 0.512 | 0.531 | 0.519 | 0.531 |
| All feature sets | 0.746 | 0.762 | 0.746 | 0.762 |

Table 10. Evaluation of the features based on Linear Regression model

| Evaluation Metrics Method | RMSE | Pearson correlation coefficient | CCR |
|---------------------------|------|---------------------------------|------|
| Temporal-textual | 1.14 | 0.26 | 0.36 |
| Global-textual | 1.11 | 0.45 | 0.39 |
| Surface | 1.13 | 0.32 | 0.32 |
| Temporal | 1.18 | 0.19 | 0.3 |
| All feature sets | 1.04 | 0.54 | 0.43 |

4.4 Evaluation of the Features

In this section, we evaluate the role of the introduced features. Table 9 shows the value of each feature set in the 3-class model. The maximum value belongs to surface and global-textual feature sets. "Category" and "Existence of featuring image" are two features that have the maximum values among surface feature set. The articles in the "politics" and "economy" categories attract more people than the other categories in Iran. Since Random Forest classifiers can combine feature sets well, the combined results are better.

Table 10 shows the value of each feature set in the Linear Regression classifier. The maximum value belongs to the global-textual feature set. The textual-temporal features have important roles in spreading news. These features extract important terms in the day of publication and the day before. These terms can be useful in various applications such as Business Intelligence Software for news agencies. Also, they can suggest interesting topics to their editors.

The textual features have a list of both unigram and bigram terms. Each terms has a score based on user favorites. Some of the highly scored terms

include name of important persons such as presidential candidate, actors, and sport teams.

5 Conclusion and Future work

The purpose of this paper was prediction of the rate of people's interest to the society issues based on user activities on web pages such as clicking on like/dislike buttons, leaving comments for articles or replying to comments. This information was extracted from news articles and comments in online news agencies. We tried to predict the VOC received for news articles before their publication.

In this paper, four feature sets were introduced that consider various factors that are effective in attracting users' interest. Two feature sets were focused on textual information. These features weight terms according to their importance. Two other feature sets were focused on surface and temporal information.

We used different classifiers and predictors to achieve the best results. To prepare the dataset, four popular online news agencies were crawled completely. In order to evaluate the proposed method, several experiments were performed. The

results revealed salient improvement in comparison with the baseline approach.

To improve the performance of the proposed method, we can consider the relation among news articles. This issue is known as topic detection and tracking (TDT)[21]. Several controversial news articles may occur together making a flood of comments. Also, we can use tweets in combination with news articles and comments. Many of the articles that appear on the news, are shared in social networks first. This information might be helpful in prediction of hot news articles. Since, the number of likes/dislikes and page visits are not available in the moment the article is published, we have to access this information 24 hours after publication. If this information is retrieved sooner, the predictor can predict hot news sooner.

References

1. **Gumbrecht, M. (2004).** *Blogs as "Protected Space"*. New York, NY, USA.
2. **Mishne, G. & Glance, N. (2006).** *Leave a Reply: An Analysis of Weblog Comments*. Edinburgh, UK, pp. 22–26.
3. **Tatar, A., Amorim M.D., Fdida, S., & Antoniadis, P. (2014).** A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, Vol. 5, No. 1, pp. 1–20.
4. **Tsagkias, M., Weerkamp, W., & Rijke, M. (2009).** Predicting the volume of comments on online news stories. *Proceedings of the 18th ACM conference on Information and knowledge management*, ACM, pp. 1765–1768.
5. **Siersdorfer, S., Chelaru, S., Nejd, W., & San Pedro, J. (2010).** How useful are your comments?: Analyzing and predicting Youtube comments and comment ratings. *Proceedings of the 19th international conference on World Wide Web*, pp. 891–900.
6. **Hsu, C.F., Khabiri, E., & Caverlee, J. (2009).** Ranking comments on the social web. *IEEE International Conference Computational Science and Engineering, CSE'09*, Vol. 4, pp. 90–97.
7. **Diakopoulos, N. & Naaman, M. (2011).** Topicality, time, and sentiment in online news comments. *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 1405–1410.
8. **Mishra, A. & Rastogi, R. (2012).** Semi-supervised correction of biased comment ratings. *Proceedings of the 21st international conference on World Wide Web*, pp. 181–190.
9. **Jamali, S., Rangwala, H., & Digging, D. (2009).** Comment Mining, Popularity Prediction, and Social Network Analysis. *Web Information Systems and Mining, WISM 2009 International Conference*, pp. 32–38.
10. **Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., Amorim, M.D., & Fdida, S. (2011).** Predicting the popularity of online articles based on user comments. *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, pp. 67.
11. **Tsagkias, M., Weerkamp, W., & De Rijke, M. (2010).** News comments: Exploring, modeling, and online prediction. *Advances in Information Retrieval*, pp. 191–203.
12. **Mishne, G. (2007).** *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam.
13. **Balali, A., Faili, H., & Asadpour, M. (2014).** A supervised approach to predict the hierarchical structure of conversation threads for comments. *The Scientific World Journal*.
14. **Chan, J., Hayes, C., & Daly, E.M. (2010).** Decomposing Discussion Forums and Boards Using User Roles. *CWSM*, Vol. 10, pp. 215–218.
15. **Hu, M., Sun, A., & Lim, E.P. (2007).** Comments-oriented blog summarization by sentence extraction. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 901–904.
16. **Breiman, L. (2001).** Random forests. *Machine learning*, Vol. 45, No. 1, pp. 5–32.
17. **Murphy, K.P. (2012).** *Machine learning: a probabilistic perspective*. The MIT Press.
18. **Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009).** The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, Vol. 11, No. 1, pp. 10–18.
19. **Schuth, A., Marx, M., & Rijke, M. (2007).** Extracting the discussion structure in comments on news-articles. *Proceedings of the 9th annual ACM international workshop on Web information and data management*, pp. 97–104.
20. **Dehghani, M., Shakery, A., Asadpour, M., & Koushkestani, A. (2013).** A learning approach for email conversation thread reconstruction. *Journal of Information Science*, Vol. 39, No. 6, pp. 846–863.
21. **Allan J. (2002).** Introduction to topic detection and tracking. *Topic detection and tracking*, Springer; pp. 1–16.

22. **Rizos, G., Papadopoulos, S., & Kompatsiaris, Y. (2016).** Predicting news popularity by mining online discussions. *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 737–742.
23. **Ren, H. & Quan, Y. (2015).** *Predicting and Evaluating the Popularity of Online News*. http://cs229.stanford.edu/proj2015/328_report.pdf.
24. **Fernandez, K., Vinagre, P., & Cortez, P. (2015).** A proactive intelligent decision support system for predicting the popularity of online news. *Portuguese Conference on Artificial Intelligence*, pp. 535–546.

*Article received on 18/12/2016; accepted on 22/02/2017.
Corresponding author is Ali Balali.*