Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 10, No. 4, November 2019), Pages: 87-96
www.jacr.iausari.ac.ir

Sentimental Categorization of Persian News Headlines Using Three Machine Learning Techniques versus Human Categorization

Vahid R. Mirzaeian✉

ELT Department, Alzahra University, Tehran, Iran
mirzaeian@alzahra.ac.ir

Received: 2020/06/07; Accepted: 2020/09/07

Abstract

The aim of this paper is to elaborate on an attempt to classify Persian news headlines using machine learning techniques rather than human-based analysis. Three major techniques were introduced and applied to Persian news headlines. Results were compared with each other as well as the human analysis. It was concluded that these techniques outperformed human analysis and one technique (Naïve Bayes) was superior to all the techniques mentioned. It can be concluded from this study that the inclusion of discourse analysis (a branch of linguistics studying sentences beyond sentence boundary) is necessary in order to attain better results since the whole sentence may convey something which is different or opposite from the elements making up the sentence. In other words, what is seen in the headline does not necessarily reflect what is mentioned in the news itself. So it is recommended that in future studies, elements from discourse analysis be introduced into these algorithms so that better results can be achieved.

Keywords: *Persian, Headlines, Sentiment, Machine Learning Method, Support Vector Machine, Maximum Entropy, Naïve Bayes*

1. Introduction

With the advent of the Internet and social networks, almost all the information particularly news is broadcasted over these digital networks. To help users locate what they are looking for, automatic text categorization techniques have been developed [1, 2, 3].

Majority of research in this filed has been devoted to topic categorization, i.e. to which subject matter the text belongs [4, 5, 6]. However, in recent years, Persian language has witnessed a rapid growth in news websites broadcasting varieties of news to Persian speakers. An important factor determining the selection of news headlines is related to the sentiment allocated to each news headlines by the reader. Here, the term 'sentiment' refers to the general positive, negative or neutral opinion associated with a news headline. Dividing news headlines into these categories will enable readers to quickly select headlines of interest.

Sentiment analysis (SA) is also useful in business intelligence applications [7,8] as well as recommender systems [9] where the system can automatically summarize user input. Indeed, using SA, one can process survey responses given in a natural language format. SA can also be used in filtering messages in order to be able to discard ‘flames’ [10, 11].

SA has broad and powerful applications. Organizations across the world adopt this feature to extract insights from social media [12]. Commercially speaking, shifts in sentiment on social media can be correlated to shifts in stock market. As a recent example, Russian government is blamed for manipulating the public media using this technique to persuade voters to vote for President Trump instead of Clinton [13].

This tool can also be an essential part of any market research and customer service approach. Using this technique, not only can marketers understand what their customers think about competitors, they can also realize what they think of the company’s products and services. Using SA, the overall experience of customers can be revealed quickly and efficiently [14].

In this paper, the applicability of three machine learning techniques in evaluation and categorization of Persian news headlines based on sentiment inherent in them will be analyzed and discussed. First, the news headlines used for the current study i.e. Hamshahri collection will be introduced. Next, the procedure employed in this study will be elaborated. The implementation of machine learning techniques in sentiment analysis will be explained in terms of Support Vector Machine, Maximum Entropy, and finally Naïve Bayes. The results will be given next in terms of unigrams, feature presence versus feature frequency, bigrams, parts of speech and location. The paper comes to an end with discussion and conclusion.

2. Review of Literature

User generated data has expanded exponentially over the Internet due to the expansion of Web 2.0. Social networking sites such as Telegram, Facebook, Instagram, Twitter, etc. offer users platforms to share their views, knowledge and experiences on recent trends in economics, politics as well as global issues. Print version of newspapers and magazines are almost scarce and majority of these media are published online [15]. Embedding the social intelligence from online news headlines is a tedious job. These problems lead to the development of analytic methods to automatically extract sentiments from news headlines known as SA.

SA collects online documents, news headlines in our case, to understand emotions, attitudes and opinions. The term ‘sentiment analysis’ was introduced by Das and Chen [16] for stock market SA. Since then, its impact can be seen in many applications such as product reviews [17], foreseeing sales and stock markets [18], and analyzing political issues [19]. Other uses include: sensitive webpage classification for content advertising [20], commonsense based intelligence system interface [21], predicting movie sales [22], prediction of negative or hostile [23], E-rule making [24] opinions on a law before approval, classification of email on the basis of emotions such as anger email, depressing email, normal mail [25] and visual SA for abstraction of subjectivity in human cognition processes [24]. SA of still images and videos is more challenging than other similar tasks. Recently, in order to conduct multimodal SA, visual features are merged with decision-level fusion methods to extract affective information from multiple modalities [23].

3. Procedure

Hamshahri collection [22] (Figure 1) was selected for the current study since it was available in digital format. However, since the data was in a raw format, data had to be hand labelled to make it appropriate for analysis. Based on a researcher [21], news headlines are very different to classify from a sentimental point of view since they tend to be highly ambiguous. The main source of data in the current study was Hamshahri collection. Hamshahri is the name of a popular newspaper in Persian and Hamshahri collection is a standard reliable Persian text collection used at Cross Language Evaluation Forum (CLEF) during years 2008 and 2009 to evaluate Persian information retrieval systems.

Figure 1. Hamshahri Corpus

Some Persian native speakers were also invited to read each headline and allocate either of the three terms (positive, negative and neutral) to them; however, in this research, the focus was on two categories, namely, positivity and negativity. Around 1000 news headlines were collected and 10 native speakers were invited to specify the sentiment.

Assigning texts to different topics can be achieved using key words; however, sentiment categorization can be easy for human beings but not that easy for machines. One way is to generate a list of words referring to the positive or negative sentiment. In order to test key word theory, two native speakers in the computation department were asked to classify some Persian news headlines into positive and negative sentiment and generate a list of key terms that can be used for this categorization. Their selected key terms have been listed in Table 1.

Table 1. Preliminary results gained from human analysts (gained from 1000 negative and 100 positive news headlines)

	Suggested key terms	**Accuracy**	**Ties**
Human analyst 1	positive key terms: خوب، عالی، زیبا، دلپذیر بد،negative key terms: وحشتناک، غاصب	59%	76%
Human analyst 2	positive key terms: محشر، صفا، وفا negative key terms: مخوف، دهشتناک	65%	40%

These key terms were later converted into a simple decision procedure which counted the number of key terms and decided if the news headline was positive or negative. We applied this procedure in Hamshahri corpus and found the accuracy to be 59% and 65% respectively. It has to be noted that the percentage of documents rated equally was not high as shown in Table 1. The data revealed that this method was working well but the accuracy level was not satisfactory. After analyzing the data, eight positive and eight negative key words were selected as shown in Table 2.

Table 2. New key words proposed after third human analysis and statistical evaluation

	new positive and negative key words	**Accuracy**	**Ties**
Human analyst 3 and statistical analysis	positive: خوب، عالی Negative: بد، وحشتناک	70%	17%

However, although the results were improved, the equality measure was quite low. Based on the data, it was discovered that it might be a good idea to use corpus based techniques in addition to key terms to improve the results. Some other corpus-based experiments were conducted with the data; however, the 70% accuracy achieved previously could not be exceeded.

3.1 Using Machine Learning Techniques

It was decided to explore if sentimental categorization could be regarded as a topic categorization or other criteria should be introduced to achieve higher accuracy. Three techniques previously used for English were tested to see how they work for Persian. These techniques were Support Vector Machine, Maximum Entropy and, Naïve Bayes categorization. In order to implement these techniques, the standard bag of feature framework was used. Each technique will be explained briefly here.

3.1.1 Support Vector Machine (SVM)

SVMs have also been proven to be highly effective in text categorization applications. They have also outperformed other algorithms such as NBs. They are considered to be non-probabilistic meaning unlike the other two mentioned algorithms, they are large-margin. In two category cases like ours, the training procedure involved finding a hyperplane represented by vector $\vec{w}$. It was both capable of separating headline vectors and setting a margin for that separation.

In this algorithm, the solution for a headline categorization can be written as in (1)

$$\vec{w} := \sum_j a_j c_j \vec{d}_j \geq 0, \qquad (1)$$

in which the a_j is obtained by solving a dual optimization problem. If $\vec{d}_j$ is greater than zero, it is called support vector since it is the only factor contributing to $\vec{w}$. In order to classify test instances, it only has to determine which side of $\vec{w}$ hyperplane it follows.

3.1.2 Maximum Entropy

ME categorization has also proven to be effective in various natural language processing (NLP) applications. Some researchers [7, 9,12] have proved that in certain conditions, it outperforms NB. In this algorithm, the estimate of P(c|d) takes the exponential form (2):

$$PME(a\,|\,d):=\frac{1}{Z(d)}\exp\left(\sum_{i}\lambda_{i,c}F_{i,c}\left(d,c\right)\right), \tag{2}$$

In which Z (d) is considered to be a normalizing function. Fi,c is considered to be a feature function for fi and class c is defined as follows (3):

$$F_{i,c}\left(d,c\right):=\left\{\frac{1,}{0}ni\left(d\right)>0andc'=c\right\} \tag{3}$$

’ or واكنش شديدFor example, a specific feature function might be initiated if the bigram ‘ ‘harsh response’ appears in the headline and the headline is deemed to be negative. It has to be noted that unlike NB, ME assumes that there is no relationship between features and as a result, it may perform better when conditional independent assumption is not met.

In this algorithm, the $\lambda'_{i,c}$ ′s are taken to be parameters carrying weight, checking the definition of PME reveals that large $\lambda_{i,c}$ is considered to be a strong indicator of class c. In order to maximize the distribution entropy, the parameter values are set subject to the fact that values of the features are equal to their expected values.

3.1.3 Naïve Bayes

An approach to text categorization is to assign class c* = arg maxc P (c |d) to a given news headline. In order to drive the NB classifier, first Bayes’ following rule should be observed (4):

$$P(c\,|\,d) = \frac{P(c)P(d\,|\,c)}{P9d)}, \tag{4}$$

in which P(d) has no role in the selection of c*. NB decomposes the term P(d|c) and assumes that fi’s are independent conditionally given d’s class (5).

$$P_{NB}(C\,|\,d) := \frac{P(c)\prod_{i=1} P(f_i\,|\,c)ni(d)}{P(d)} \tag{5}$$

In our training method, the relative frequency of p(c) and p (fi | c) were estimated using add-on smoothing.

NB-based text categorization has proven to work surprisingly well although it is very simple and its conditional independence assumption does not take place in real-word situations. Some researchers [31, 32] have shown that NB is very effective in certain situations where abundant dependent features are available. Since more complicated algorithms may provide better results, two such algorithms will be explored in the following sections.

4. Results

As mentioned in the previous section, Hamshahri corpus was implemented to test these three techniques. In order to create a dataset, with uniform class distribution, around1000 positive and 1000 negative sentiment headlines were randomly selected. Later, each was randomly divided into three equal groups.

In order to perform the test, all possible punctuation marks such as commas and hyphens were removed from the headlines. one unorthodox method was also implemented to have better results. Headlines with covert signs relating to negative or positive sentiment were deleted so words like ‘خوب’ meaning ‘good’ and ‘بد’ meaning ‘bad’ were removed from our corpus.

In this study, the focus was mainly on features based on bigrams and unigrams. Since training ME has proven to be expensive and time-consuming, the main focus was on 1) those unigrams appearing at least five times in our corpus and 2) bigrams occurring most often in the data.

Unigrams

Sentimental categorization limited to unigrams have been displayed in line 1 of Table 3. It is clearly evident that machine learning algorithms achieved beyond 50%. Moreover, they outperformed human analysts’ performances of 59% and 65%.

Table 3. Results of three techniques shown in percentages. The best performance has been highlighted.

	Feature	Number of Features	Presence or Frequency	Support Vector Machine	Maximum Entropy	Naïve Bayes
1	Unigram	16166	Frequency	72.9	Not available	**78.8**
2	Unigram	16166	Presence	**82.8**	80.5	81.1
3	bigram and unigram	32332	Presence	**82.8**	80.9	80.7
4	bigram	16166	Presence	77.2	**77.5**	77.4
5	parts of speech and unigram	16696	Presence	**81.9**	80.5	81.6
6	adjective	2634	Presence	75.2	**77.7**	77.1
7	top 2634 unigrams	2634	Presence	**81.5**	81.1	80.4
8	location and unigram	22431	Presence	**81.7**	80.2	81.1

It has to be noted that some researchers have reported that these algorithms perform up to 90% or more in topic-based categorizations. Although the results were satisfactory, ways were sought to improve such performance.

Feature presence versus feature frequency

In ME feature functions, only the presence of features is taken into account. In order to make sure if reliance on frequency is a good criterion, the vectors were converted into binary and both NB and SVM were reran on the new vectors.

As line 2 in the table indicates, by focusing on feature presence not feature frequency, better results were achieved. It is interesting to note that this is opposite to what some researchers [20] have found regarding topic categorization of texts. This reminds us of the fact that news headlines are drastically different from documents and the differences should be taken into account in future studies.

4.1 Bigrams

In addition to unigrams and feature frequency, bigrams were also taken to consideration to find out how they helped to determine the sentiment of the news headlines. Line 3 in Table 3 is related to the information regarding bigrams. It has to be noted that based on Pederson [19] bigrams per se can be a good measure of word sense disambiguation.

Comparing line 2 and 4 in Table 3, it was realized that reliance on bigrams only could negatively affect the results. So it was concluded from this result that bigrams should be used in conjunction with other features to get an acceptable result.

4.2 Parts of Speech (POS)

It was also decided to add POS tags to the data to investigate if this can also affect the results. As line 5 in Table 3 indicates, this tag insertion is not big enough to justify its use given the time and energy it takes to add tags to the data.

Since some researchers have focused on adjectives to detect sentiments in documents, it was decided to test this method as well. Contrary to our expectations that adjectives are normally biased and can carry a lot of information regarding the writer's attitude toward a topic, the data as depicted in line 6 of Table 3 revealed that it was not changing the results drastically as expected. Line 7 in Table 3 clearly indicates that using the most frequent unigrams is useful and sufficient in determining the sentiment of Persian news headlines.

4.3 Location

It was also hypothesized that the location of certain vocabulary items in the headline could also have an effect on sentiment detection. The data were tagged with labels such as INI (initial), MED (Media) and FIN (final) to refer to the position of key terms in the headline. Results in line 8 of Table 3 indicated that this was not satisfactory and the location could be safely ignored in the analysis without serious effect on the results.

5. Discussion

Majority of research in the field of sentiment categorization has been based on knowledge and intuition. Our work here has been focusing on semantic categorization based on some specific key terms with the implementation of pre-selected seed words [23, 24, 25]. Other methods based on the determination of the whole text have mainly been based on cognitive modes presented by cognitive linguists [18, 19, 20]. Our study revealed that human being are not necessarily successful in determining the sentiment of a headline by merely identifying some key terms.

Results as shown in table 3 clearly indicated that these techniques could outperform human analysis. The data also revealed that although Naïve Bayes tended to be the worst and SVM to be the best, their differences were negligible.

In spite of using several features as shown in table 3, we failed to gain the degree of accuracy compared to topic categorization of documents. This can partially be due to the fact that news headlines tend to be ambiguous to attract the attention of readers. It was concluded that unigram feature detection was proven to be the most successful. In fact, the difference was so little that all the other features could easily be ignored without affecting our final result. As mentioned before, it is interesting to note that unlike topic categorization in which frequency information plays a significant role [10, 11], in our study, presence information played a major role in sentiment detection and categorization.

In order to find out why this difference was observed, we had to review the data once more and come up with some explanation. As mentioned before, words carry different meanings and they reveal the meaning within the context. Since we focused our attention to the news headlines only, and did not take into account the news itself, it was quite hard to decide if a news headline was positive or negative. In addition, journalists are constantly advised not to be biased in their reports. This puts a pressure on them to find ambiguous items and use them in their headlines to disguise their bias toward a piece of news. We showed some examples to some native Persian speakers but they could not reach a general agreement whether the headlines were positive or negative.

6. Conclusion

It can be concluded from this study that the inclusion of discourse analysis is necessary in order to attain better results. Some researchers assert that the whole is not necessarily the sum of the parts. It means that what you see in the headline does not necessarily reflect what is mentioned in the news itself. As for the future studies, it is recommended that elements from discourse analysis be introduced into these algorithms so that better results can be achieved.

References

[1] Farhad Rezvani; Farhad Soleimanian Gharehchopogh. "A Novel Approach to Feature Selection Using PageRank algorithm for Web Page Classification". Journal of Advances in Computer Research, 10, 4, 2019.

[2] Thippa Reddy, Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava and Thar Baker. Analysis of dimensionality reduction techniques on big data. IEEE Access, 8, 54776-54788, 2020.

[3] Rajesh Kaluri and Pradeep Reddy. A framework for sign gesture recognition using improved genetic algorithm and adaptive filter. Cogent Engineering, 3(1), 1251730, 2016.

[4] Praveen Kumar Reddy, Maddikunta,Thippa Reddy Gadekallu, Rajesh Kaluri, Gautam Srivastava, Reza Parizi and Mohammad Khane. Green communication in IoT networks using a hybrid optimization algorithm. Computer Communications. Computer Communications, 159, 97-107, 2020.

[5] Rajesh Kaluri & Pradeep Reddy. Sign gesture recognition using modified region growing algorithm and adaptive genetic fuzzy classifier. International Journal of Intelligent Engineering and Systems, 9, 225-233, 2016.

[6] Masrour Dowlatabadi; Ahmad Afshar; Ali Moarefianpour. "Simultaneous Classification and Traction of Moving Obstacles by LIDAR and Camera Using Bayesian Algorithm". Journal of Advances in Computer Research, 10, 4, 2019.

[7] Teymour Shahi and Keneth Pant. Nepali news classification using Naïve-Bayes, support vector machines and neural networks. in 2018 International Conference on Communication, Information Computing Technology (ICCICT), Feb. 2–3, Mumbai, India, 2018.

[8] Wendy Jirasirilerd and Peter Tangtisanon. Automatic labeling for Thai news articles based on vector representation of documents, in 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST), 2018.

[9] Pardeep Kaushik and Amar Sharma. Literature survey of statistical, deep and reinforcement learning in natural language processing. in International Conference on Computing, Communication and Automation (ICCCA), 2017.

[10] Chang Li, Gary Zhan and Zing Li. News text classification based on improved Bi-LSTM-CNN. in 9th International Conference on Information Technology in Medicine and Education (ITME), 2018.

[11] Jing Zhang, Yori Li, Jang Tian and Tashu Li. LSTM-CNN hybrid model for text classification. in IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2018.

[12] Shiro Sharma, Jing Agrawal, Senehi Agrawal and Suran Sharma. Machine Learning Techniques for Data Mining: A Survey (IEEE), 2013.

[13] Zeinab Madhoushi, AhmadHamdan and Sepehr Zainudin. Sentiment analysis techniques in recent works, in Science and Information Conference, (IEEE), 2015.

[14] Somayyeh Jafarali Jassbi; Farzaneh Jahanshahi Javaran; Hossein Khademolhosseini; Amir Sabbagh Molahosseini. "Design and Analysis of a Fault Tolerant 3-Input Majority Gate in Quantum-dot Cellular Automata". Journal of Advances in Computer Research, 10, 4, 2019.

[15] Ted Bui, Daniel Nguyen and Tim Ngo. Supervising an unsupervised neural network. In First Asian Conference on Intelligent Information and Database System (IEEE), 2009.

[16] Dang Cecchini and Li Na. Chinese news classification. In IEEE International Conference on Big Data and Smart Computing, 2018.

[17] Uma Mohiuddin, Hanady Ahmed and Madi Ismail. NEWSD: a real time news classification engine for web streaming data. in International Conference on Recent Advances in Computer Systems (RACS) , 2015.

[18] Kumar Gurmeet and Beau Karan. News classification and its techniques: a review. IOSR J. Comput. Eng. (IOSR-JCE) 18(1) Ver. III, 2016.

[19] Valery Rao and Jim Sachdev. A machine learning approach to classify news articles based on location. in Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS), 2017.

[20] Jim Azzopardi and Chris Staff. Fusion of news reports using surface-based methods. in 26th International Conference on Advanced Information Networking and Applications Workshops, 2012.

[21] Jang Zhang, Chen Lu, Mung Zhou, Sin Xie, Yuri Chang and PengYu. HEER: heterogeneous graph embedding for emerging relation detection from news. In IEEE International Conference on Big Data (Big Data), 2016.

[22] Dutu Nagalavi and Madi Hanumanthappa. A new graph based sequence clustering approach for news article retrieval system. in IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017.

[23] Mike Uhl. Explaining U.S. consumer behavior with news sentiment. ACM Trans. Manag. Inf. Syst. 2, Article 9, 2011.

[24] Cross Mason, Buris McInnis and Sim Dalal. Machine learning for the automatic identification of terrorist incidents in Worldwide News Media. In 2012 IEEE, ISI 2012 June, 11–14, Washington D.C., USA, 2012.

[25] Abolfazl AleAhmad , Hadi Amiri , Ehsan Darrudi , Masoud Rahgozar , Farhad Oroumchian, Hamshahri: A standard Persian text collection, Journal of Knowledge-Based Systems, Vol. 22 No.5, p.382-387, Elsevier, 2009.

[26] Valery Bobichev and Omar Kanishcheva. Sentiment analysis in the Ukrainian and Russian news. In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017.

[27] Buris Ko, Chris Park, Daniel Lee, Jimmy Kim, Harold Choi and Danile Han. Finding news articles related to posts in social media: the need to consider emotion as a feature. In 2018 IEEE International conference on Big Data, 2018.

[28] Gary Abali, Eli Karaarslan, Alice Hurriyetoglu and Francis Dalkilic. Detecting Citizen problems and their locations using Twitter data. In 6th International Istanbul Smart Grids and Cities Congress and Fair (ICSG) (IEEE), 2018.

[29] Nisha Kaur. A Survey of Clustering Techniques and Algorithm (IEEE), 2015.