Developing a machine learning model for fake news detection

Rodion Filippov*, *Anna* Sazonova, and *Yuri* Leonov

Bryansk State Technical University, Bryansk, 241013, Russia

Abstract. The article is devoted to the problem of detecting fake news. This issue is relevant nowadays. Fake news paves the way for deceiving people and promoting ideologies. People who provide incorrect information benefit by earning money from the number of interactions with their publications. One of the typical tasks that arise in the process of identifying news is determining whether news belongs to one of two classes, namely the fake news or the real news. With the help of modern methods of machine learning and primary data processing, this problem is effectively solved.

1 Introduction

Fake news hurts businesses as any industry can be targeted by disinformation companies. Modern text classification technologies based on machine learning make it possible to speed up and partially automate the process of filtering out the fakes.

The concept of "fake news" has a number of definitions, the common feature of which is the intention to mislead and spread false information on what is happening in the world to obtain some benefit or plant certain ideas and opinions. Fake news can be presented in the form of publications under enticing "clickbait" headlines, in the form of propaganda, "the author's opinion", humor and satire, etc.

One of the typical tasks that arise in the process of identifying news is determining whether the news belongs to one of two classes, namely the fake news or the real news. With the help of modern methods of machine learning and primary data processing, this problem is effectively solved.

Today, many hopes are associated with machine learning (ML), but its application success is determined not only by the algorithm choice which is adequate to the task, but also by the right steps at the stages of planning, developing and implementing the model.

Each stage of building machine learning has a big impact on the final result. Further, the steps will be discussed in more detail.

2 Method

It is very important to collect data correctly, as these are the quality and quantity of data that affect how good the model is and how precise the desired result will be [1].

* Corresponding author: libv88@mail.ru

The high quality of the data allows obtaining more precise results.

However, it is significant to strike a balance here. Not always a huge amount of data is an advantage. As a general rule, the larger the data set is, the more difficult it is to use it correctly and get the right findings.

For the study, a data set is taken from the Kaggle website. Kaggle is a data mining competition system and social network for data scientists and machine learning experts. Kaggle is a subsidiary of Google LLC.

The environment is organized as a public web platform where users and organizations can publish datasets, explore and build models, work with other data scientists and machine learning engineers, arrange and enter competitions to solve data science challenges. The system hosts open data sets, provides cloud-based tools for data processing and machine learning. [2, 3]

Link to view the data set is https://www.kaggle.com/c/fake-news/data.

The site provides two sets of data, namely a training set and a test set.

For the fake news detection task, the training data set contains the following attributes:

- the title of the news article;
- the author of a news article;
- the text of the article (may be incomplete);
- the label which marks the article as potentially untrustworthy (1 means untrustworthy, 0 means trustworthy).

The training dataset contains 20800 records and 6 columns.

The test dataset contains 5200 records and 5 columns (the “label” column is not included).

Choosing the right type of machine learning depends on several factors, including the data size, its quality and variety, and understanding what answers based on this data are needed to solve the problem. Also, attention should be paid to precision, training time, parameters, data. Therefore, choosing the right algorithm is a combination of business needs, specifications, experimental work, and consideration of available time.

Even the most experienced data scientists can’t say which algorithm will work best without experimenting with different kinds of algorithms.

After going through the scheme for choosing a machine learning model, it is revealed that the task of detecting fake news is a classification one.

The task of classification is a formalized one in which there are many objects or situations which are in some way divided into classes. A finite set of objects is given for which it is known which classes they belong to. This set is called a sample. The class belonging of the rest of the objects is not known. It is required to build an algorithm capable of classifying an arbitrary object from the initial set [5,6].

Mathematical statement of the problem is the following: Let X be the set of descriptions of objects, Y be the set of numbers or names of classes. There is unknown target dependence, namely mapping $y^*: X \to Y$, the values of which are known only on the objects of the final training sample:

$$X^m = \{(x_1, y_{1,\cdots,}(x_m, y_m))\} \quad (1)$$

It is required to construct an algorithm $a: X \to Y$, capable of classifying an arbitrary object $x \in X$.

The next step is to choose the right model. There are a huge variety of models: some of them are better suited for working with pictures, some are better adapted for working with music, some of them are better able to deal with a set of numbers.

Consider the main methods of machine learning for solving the classification problem.

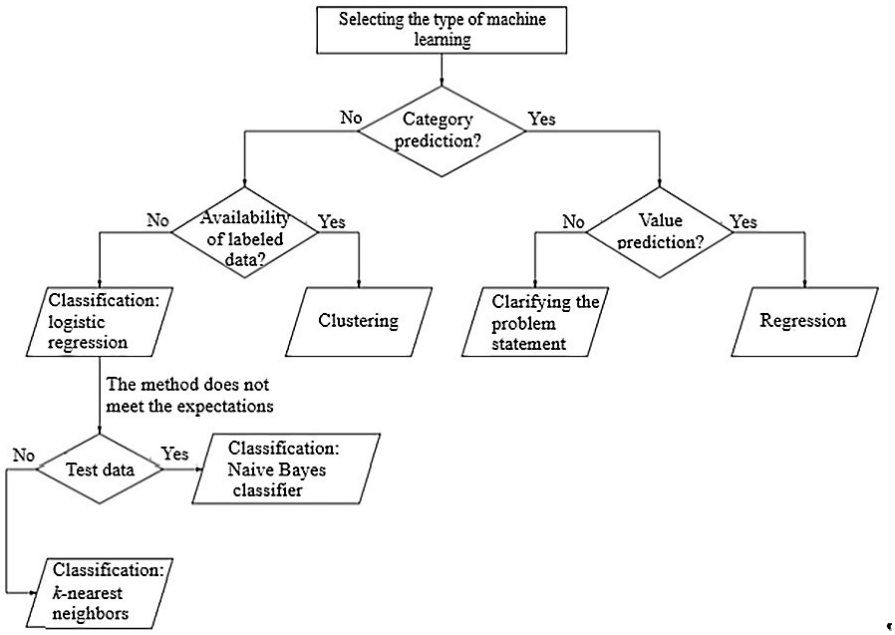

Fig. 1. Flowchart for choosing the type of machine learning [Compiled by the authors].

Naive Bayes classifier. Let objects be described by n -dimensional feature vectors, $x = (x^1, \ldots, x^n)$, where each feature has its own range of acceptable values, $x^j \in D_j$. When the sets D_j are different, the features are said to be heterogeneous. In the applied problems of classification and pattern recognition, the most common cases are real, integer and binary features [6,7].

Logistic regression. Logistic regression is a method for constructing a linear classifier that allows estimating posteriori probabilities that objects belong to classes.

Support vector machine. One of the most popular machine learning methods, the Support Vector Machine (SVM) is developing the ideas proposed by V.N. Vapnik and A.Ya. Chervonenkis during 1960-1970. The method took its final shape in 1995, when it was shown how kernels could be effectively used in this method [7,8].

Decision tree. The main idea of decision trees is to recursively part the feature space by splits, which are hyperplanes parallel to coordinate hyperplanes (if the feature is quantitative) or by categories (if the feature is nominal).

KNN (K-nearest neighbors). The k -nearest neighbor method is one of the simplest and at the same time universal methods used both for solving classification problems and regression recovery. In the case of classification, a new object is classified by assigning it to the class that is predominant among the k nearest (in the feature space) objects from the training set. If $k = 1$, then the new object belongs to the same class as the nearest object from the training sample [8,9].

In some cases, these responses are taken into account with weights that are inversely proportional to the distance to the object [10, 11]. This is especially useful for solving a classification problem with unbalanced data, i.e. when the number of objects belonging to various classes is very different [12].

When choosing a model, you need to consider the following parameters:

- Accuracy.
- Training time.
- Linearity.

– Number of parameters.

Comparison of methods according to these criteria is presented in the table:

Table 1. Comparative characteristics of the machine learning methods [Compiled by 13,14].

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Logistic regression	-	•	•	5	
KNN	•	-	-	9	Additional customization is possible
Support vector machine	-	○	•	5	It is good for large feature sets
Decision tree	•	○	-	6	
Naive Bayes classifier	-	○	•	3	

Notes: • — Demonstrates excellent accuracy, short training time and use of linearity. ○ — Shows excellent accuracy and average training time.

According to this comparative characteristic, it follows that the method of logistic regression is the leading one, which shows the use of linearity and a short training time. Therefore, when benchmarking against the main criteria, the best machine learning method is logistic regression [15,16].

3 Results

These stages are given for simulating the problem solution of detecting fake news.

Stage 1. Reading a data set from a file.

– connecting necessary libraries;
– loading the data set;
– data preview.

Stage 2. Data pre-processing.

– checking for zero values in training and testing datasets. If there are such entries, you must replace NaN values with spaces;
– uniting the columns “title” and “author”;
– splitting the data by the “label” attribute (1 means fake news, 0 means real news);
– having the necessity to transform the data to a productive form to feed the machine learning model as input:
    1. All sequences, except for English characters, are removed from the string.
    2. All characters in strings are converted to lowercase to avoid false predictions or ambiguities with the uppercase and lowercase letters.
    3. All sentences are tokenized into words. Tokenization involves separating sentences and words from the body of text.
    4. Stemming is applied to tokenized words to facilitate quick processing.
    5. The words are then combined and stored in the corpus.

Stage 3. Dividing the data set into training, testing and validation sets. At this stage, the train_test_split technique is used, which splits the training, testing and validation data into a ratio: 80/10/10.

Stage 4. Creating a machine learning model.

In this step, a machine learning model is created and scored against various metrics shown using the classification report. Implementing logistic regression is given (Figure 2).

	precision	recall	f1-score	support
	0.50	0.09	0.15	11
0.0	0.96	0.87	0.91	1040
1.0	0.87	0.96	0.91	1029
accuracy			0.91	2080
macro avg	0.78	0.64	0.66	2080
weighted avg	0.91	0.91	0.91	2080

Fig. 2. Classification report [Compiled by the authors].

The accuracy of the machine learning model is 0.91.

One of the main characteristics of the classifier is its accuracy. The most obvious way to evaluate accuracy is the percentage of correct answers.

$$Accuracy = \frac{P}{N}, \tag{2}$$

where P is the number of objects whose class matched the class determined by the algorithm, and N is the size of the test set.

Precision and recall are metrics that are used in evaluating most information extraction algorithms [17].

To define them, some concepts are introduced:

If the classifier identified that the object class is 0 and its class is really 0, it is True Negative (TN), i.e. the classifier correctly predicted the class as negative.

If the classifier identified that the object class is 0 and its class is 1, it is False Negative (FN), i.e. the classifier incorrectly predicted the class as negative.

If the classifier identified that the object class is 1 and its class is 0, it is False Positive (FP), i.e. the classifier incorrectly predicted the class as positive.

If the classifier identified that the object class is 1 and its class is 1, it is True Positive (TP), i.e. the classifier correctly predicted the class as positive.

$$TP + FN = All\ are\ positive, \tag{3}$$

$$TN + FP = All\ are\ negative, \tag{4}$$

The precision is the ratio of the number of correctly identified positive objects to the total number of positive objects that the classifier assigned to this class, that is, the precision of determining positive answers. The closer the precision is to one, the fewer incorrect class definitions that are considered correct.

$$Precision = \frac{TP}{(TP + FP)}, \tag{5}$$

The recall is the ratio of the objects belonging to the class found by the classifier to the total number of positive objects in the test set. This measure shows how well positive responses are predicted from all positive answers in the set during the classification.

$$Recall = \frac{TP}{(TP + FN)}, \tag{6}$$

The importance of precision and recall depends on the specific task. Accordingly, there is a function that makes preferences for recall or precision, the so-called F -measure:

$$F = (b^2 + 1)\frac{Precision * Recall}{b^2 Precision * Recall}. \tag{7}$$

ROC (receiver operating characteristic, error curve) is a curve that allows evaluating the quality of a binary classification, it is set parametrically:

$$x = \frac{FP}{TN+FP}, \tag{8}$$

$$y = \frac{TP}{TP+FN}. \tag{9}$$

x is False Positive Rate, that is, the ratio of incorrectly identified elements into the positive class to all the elements of the negative class.
y is True Positive Rate, that is, the ratio of correctly identified elements in the positive class to all the elements of the positive class.

The quality of the classifier is determined by AUC which is the area under the curve, and is a quantitative characteristic for the ROC curve:

$$AUC = \int_0^1 \frac{TP}{TP+FN} d\frac{FP}{TN+FP}, \quad (10)$$

To test a machine learning model, it is necessary to make a prediction on a test data set and then compare the obtained values, i.e. predicted values with the test values (Figure 3).

	y_test	y_pred
0	1	1
1	1	0
2	0	1
3	1	1
4	0	1
5	0	0
6	1	0
7	1	1
8	1	1
9	1	1

Fig. 3. Python code snippet for comparing the obtained values [Compiled by the authors].

The comparison of the test and predicted values is given below.

This table shows that the machine learning model works correctly in detecting fake news, as it achieved an accuracy of 91%.

4 Discussion

An important step in writing the article was to choose a machine learning model. When performing this stage, it was revealed that the task of detecting fake news belonged to the tasks of classification. This stage helped to create the correct mathematical model. The next key point was to study the main methods of machine learning for this type of problem. Next, a comparative analysis of the methods was carried out according to the main criteria (accuracy, training time, linearity and number of parameters). After this step, it was decided to implement linear regression as a machine learning method for solving the problem in Python. The final step was to check the correctness of the model choice by implementing all the methods in Python and to visually present the results on the charts.

5 Conclusion

In this article, the problem of detecting fake news was considered. The aim of this article was to build the most accurate model that will correctly answer the question in most cases.

References

1. T. Graepel, J.Candela, T. Borchert, R. Herbrich*, Web-Scale Bayesian click-through rateprediction for sponsored search advertising in Microsoft's Bing search engine,* in Proceedings of 27th International Conference on Machine Learning, 13-20 (2010)
2. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, 2014)
3. C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006)
4. L.P. Coelho, V. Richard, *Building machine learning systems in Python* (Packt Publishing, 2016)
5. T. Bayes, *An essay, towards solving a problem in the doctrine of chances* (Philos Trans RSoc Lond, 1763)
6. R. Pirmagomedov, D. Moltchanov, A. Ometov, Kh. Muhammad, S. Andree, Ye. Koucheryavy, IEEE Access **7**, 180700-180712 (2019)
7. J. Demsar, Journal of Machine Learning Research **7**, 1-30 (2006)
8. A. Smola, B. Scholkopf, Statistics and Computing **14**, 199-222 (2004)
9. Xu Rui, D. Wunsch, IEEE Transactions on neural networks **16(3)**, 645-678 (2005)
10. M. Chen, Sh. Mao, Yu. Mobile, Networks and Applications **19**, 171-209 (2014)
11. Ya. LeCun, Yo. Bengio, G. Hinton, Nature **521(7553)**, 436-444 (2015)
12. A.A. Kuzmenko, D.E. Kondrashin, Ergodesign **4(6),** 230-240 (2019)
13. A.A. Kuzmenko, A.V. Averchenkov, A.S. Sazonova, *Neural network analysis of ecological and floral classification as a basis for protection of regional biodiversity*, in the collection: IOP Conference Series: Materials Science and Engineering **753**, 042029 (2020)
14. A.A. Kuzmenko, S. Kondratenko, K. Dergachev and V. Spasennikov, *Ergonomic support for logo development based on deep learning,* in the proceedings: PROCEEDINGS of the seminar CEUR 30. Ser. "Graphics in 2020 - Proceedings of the 30th International Conference on Computer Graphics and Machine Vision" **2744** (2020)
15. D.R. Kalugin, Yu.A. Leonov, RA. Filippov, L.B. Filippova, *Development of an information and analytical system for modeling the demographic situation in the russian federation,* in the proceedings: III International Workshop on Modeling, Information Processing and Computing (MIP: Computing-2021), CEUR Workshop Proceedings **2899**, 133-140 (2021)
16. A.V. Ivanova, L.B. Filippova, *Methods of text processing and machine learning when creating chatbots*, in the proceedings: New Horizons. VIII scientific and practical conference with international participation. Collection of materials and reports, Bryansk, 289-292 (2021)
17. D.I. Kopeliovich, R.A. Filippov, L.B. Filippova, E.O. Trubakov, *Theoretical and methodological support for monitoring socio-economic systems using data warehouses in OLAP technology: monograph* (Direct-Media, Moscow, Berlin, 2021)