Mikhail O. Granik, Vladimir I. Mesyura
Vinnytsia National Technical University, Vinnytsia, Ukraine
mesyura@vntu.edu.ua

DOI: 10.25045/jpit.v09.i2.06

FAKE STATEMENTS DETECTION WITH ENSEMBLE OF MACHINE LEARNING ALGORITHMS

The paper is devoted to an attempt of classifying statements made by public figures as true or false (fake). It is suggested to use number of different machine learning techniques for that and uniting them to a single system (ensemble) which predicts probability that given statement is true or not and performs the appropriate classification.

Keywords: fake news, fake statements, machine learning, deep learning, ensemble.

Introduction

Internet substantially extended the possibilities for its users to find the news that they are interested in. The progress in modern information technologies brings us to the era in which information is as accessible as ever. It is possible to find answers to the questions we are interested in in a matter of seconds. Availability of mobile devices makes it even more convenient for users. The access to the news information appears almost instantly after the event happens. We can receive updates using any device with Internet access.

This factor changed the way of getting information. Every mainstream mass media has its own online portal, Facebook account, Twitter account etc., so people can access news information really quickly.

Together with such advantages of the current state of the Internet, new challenges also emerge. Unfortunately, the news information that we get is not always true. Paradoxically, the Internet makes it harder to fact-check the available information, because there are too many sources that often even contradict each other.

Mass media generates a huge impact on the society, and always there are people who want to take advantage of it. There even exist lots of websites that produce fake news almost exclusively. They deliberately publish hoaxes, propaganda and disinformation purporting to be real news – often using social media to drive web traffic and amplify their effect. The main goal of fake news websites is to affect the public opinion on certain matters (mostly political). Examples of such websites may be found in Ukraine, United States of America, Germany, China, and many of other countries [1]. Thus, fake news is a global issue and challenge.

Many scientists believe that fake news issue may be addressed by means of machine learning and artificial intelligence [2]. There is a reason for that: recently, artificial intelligence algorithms have started to work much better on lots of classification problems (image recognition, voice detection and so on), because hardware is cheaper, and bigger datasets are available. The rise of deep learning and other artificial intelligence techniques showed us that they could be very effective in solving complex, sometimes even non-formal classification tasks.

This article describes a method for classification of short political statements by means of artificial intelligence, specifically using ensemble of such techniques. Several approaches were implemented, united in a single system and tested on a data set of a statements made by real-life politicians.

Description of the data set used for training and testing

The data set that was used for training and testing was collected by a RAMP studio team [3]. It contains of short statements made by famous public figures. Six possible labels were available for the statement. They are:

a) ‘Pants on Fire!' (completely false)
b) 'False'

c) 'Mostly False'
d) 'Half-True'
e) 'Mostly True'
f) 'True'

Each entry in the data set, besides the statement itself, also contains a lot of metadata. It contains the date when the statement was made, the job position of the public figure who made that statement, the source where the statement was taken from, some keywords that characterize the content of the statement and many more other features. The data set consists of 10460 entries in total (7569 of them were provided for training and 2891 for testing). There are more than 2000 different sources of the statements. The RAMP studio team collected the data set using PolitiFact website.

The PolitiFact is a project operated by Tampa Bay Times in which reporters from the Times and affiliated media fact-check statements by members of the United States Congress, the White House, lobbyists and interests groups. They publish original statements and their evaluations on the PolitiFact.com website, and assign each a "Truth-O-Meter" rating [4]. PolitiFact.com was awarded the Pulitzer Prize for National Reporting in 2009 for "its fact-checking initiative during the 2008 presidential campaign that used probing reporters and

the power of the World Wide Web to examine more than 750 political claims, separating rhetoric from truth to enlighten voters". At some points, PolitiFact was criticized by both liberal and conservative wings of American politics, but nevertheless it is a viable source of fact-checked information. This makes a data set useful for creating a system which will classify statements as true or false.

Data preprocessing

Prior to actual application of the artificial intelligence algorithms to the data, it should be pre-processed [5].

First of all it was decided to use only the statements themselves for classification purposes. This means that none of the metadata provided is used for classification. The classification algorithm might actually be improved in the future by taking into account this metadata.

The steps that were used for the pre-processing are the following:

a) Splitting the statements into separate tokens (words).
b) Removing all numbers.
c) Removing all punctuation marks.
d) Removing all other non-alpha characters
e) Applying the stemming procedure to the rest of the tokens. In linguistic morphology and information retrieval, stemming (or lemmatization) is the process of reducing inflected or derived words to their word stem, base or root form – generally, a written word form [6]. This helps to treat similar words (such as “write” and “writing”) as the same words and might be extremely helpful for classification purposes.
f) Removing stop words. Stop words are the words occur in basically all types of texts. These words are common and they do not really affect the meaning of the textual information, thus it might be useful to get rid of them [7].
g) Substitution of words with their tf-idf scores. In information retrieval, tf–idf, which is a shorten version of “term frequency–inverse document frequency”, is a numerical statistic measure reflects the importance of a certain word to a document in a collection or corpus [8]. The tf-idf value increases proportionally to the number of times a word appears in the document and decreases proportionally to the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. According to tf-idf, the weight of a term that occurs in a document is proportional to its frequency, and the specificity of a term can be calculated as an inverse function of the number of documents that contain the specified term.

Ensembles

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance that could be from any of the constituent learning algorithms alone. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble consists of only a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives [9].

Supervised learning algorithms are most commonly described as performing the task of searching through a hypothesis space to find a suitable hypothesis that will make good predictions with a particular problem. Even if the hypothesis space contains hypotheses that are very well-suited for a particular problem, it may be very difficult to find a good one. Ensembles combine multiple hypotheses to form a (hopefully) better hypothesis. The term ensemble is usually reserved for methods that generate multiple hypotheses using the same base learner. The broader term of multiple classifier systems also covers hybridization of hypotheses that are not induced by the same base learner [9].

Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, thus ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. Fast algorithms such as decision trees are commonly used in ensemble methods (for example Random Forest), although slower algorithms can benefit from ensemble techniques as well [9].

By analogy, ensemble techniques have been used also in unsupervised learning scenarios, for example in consensus clustering or in anomaly detection [9].

Empirically, ensembles tend to yield better results when there is a significant diversity among the models. Many ensemble methods, therefore, seek to promote diversity among the models they combine. Although perhaps non-intuitive, more random algorithms (like random decision trees) can be used to produce a stronger ensemble than very deliberate algorithms (like entropy-reducing decision trees). Using a variety of strong learning algorithms, however, has been shown to be more effective than using techniques that attempt to dumb-down the models in order to promote diversity [9].

General description of the Ensemble system

The data set was initially split into three subsets: training, validation (used for metaparameters tuning) and testing (used for getting the unbiased estimate of how well an algorithm performs on the previously unseen data).

Several machine learning techniques were implemented for the classification of short statements. Among them, there were such techniques as logistic regression, naive Bayes classifier, Random Forest classifier, Support Vector Machines and Deep Neural Networks. The results of each of this methods used separately can be seen in [5].

It was decided to use stacking ensemble for combining the results.

Stacking (sometimes called stacked generalization) involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is used to make a final prediction using all the predictions of the other algorithms as additional. Stacking typically yields performance better than any single one of the trained model. [9]

Several combining algorithms were used to make the final classification decision. They are:

a) Simple voting. Each model casts exactly one vote for the class to which a statement belong according to this model. The votes received from each model are added. The class that has the most votes in proclaimed as a result of classification. In case of equality the classification result is selected randomly between the classes with most number of votes.

b) Weighted voting. Once again each model casts exactly one vote for the class to which a statement belong according to this model. The vote of the model than is weighted according to classification accuracy of the model on the validation data set. The weighted votes received from each model are added. The class that has largest sum of the weighted votes is proclaimed as a result of classification. In case of equality the classification result is selected randomly between the classes with most number of votes.

Several sets of models were used for stacking. All of the subsets of the trained model with size bigger than or equal to three were stacked into an ensemble.

For all of the ensembles two different metrics were measured:

a) Classification accuracy based on six available categories

b) Binary classification accuracy. This metric counts the accuracy as if there were only 2 possible categories for the statement – true (based on the last three categories described above) and false (based on the first three categories described above)

The results of each stacking set can be viewed in the Table 1. Some results were omitted from the table because there results were not good enough or not informative enough.

Table 1

The results of each stacking set

Algorithm used in the ensemble	Simple voting		Weighted voting	
	Classification accuracy	Binary classification accuracy	Classification accuracy	Binary classification accuracy
Logistic regression, Naive Bayes classifier, Random Forest classifier	79%	83%	79%	84%
Random forest classifier, Support Vector Machines, Deep Neural Network	81%	85%	82%	86%
Random forest classifier, support vector machines, deep neural network, logistic regression	81%	85%	80%	85%
Random Forest classifier, Support Vector Machines, Deep Neural Network, Naive Bayes classifier	81%	88%	82%	88%
Random Forest classifier, Support Vector Machines, Deep Neural Network, Naive Bayes classifier, Logistic regression	80%	86%	81%	86%

As one can see, the results are generally improved in comparison to [5], which implies that using ensemble learning indeed improves the performance of a system.

The best results among all of the ensembles showed the one that uses such algorithms as Random Forest classifier, Support Vector Machines, Deep Neural Network, Naive Bayes classifier. The ensemble consisting of all five algorithms showed worse results – probably, because of low classification accuracy of logistic regression algorithm.

Conclusion

In this paper, several algorithms for classifying statements made by public figures were implemented and united into a single ensemble system.

The best results were shown by the stacked ensemble of the following algorithms: random forest classifier, support vector machines, deep neural network, naive Bayes classifier.

Achieved results might be significantly improved. It is possible to both improve the data which is used for training as well as the machine learning models themselves. We suggest the following possible improvements:

a) include metadata to the training process;
b) get more data and use it for training;
c) investigate misclassified examples;
d) try other machine learning approaches [10, 11].

Together with the text summarization (the problem that also can be solved by means of artificial intelligence), the approach, described in the paper, might be used for classification of news articles as fake or true. This might be a subject for future research.

References

1. Granik M., Mesyura V. Fake news detection using naive Bayes classifier / 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp.900–903.
2. Metz C. The bittersweet sweepstakes to build an AI that destroys fake news www.wired.com/2016/12/bittersweet-sweepstakes-build-ai-destroys-fake-news/
3. Fake news RAMP: classify statements of public figures. www.ramp.studio/problems/fake_news
4. The Principles of the Truth-O-Meter: PolitiFact’s methodology for independent fact-checking. www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/. Accessed Mar. 24, 2018.
5. Granik M., Mesyura V., Yarovyi A. Determining fake statements made by public figures by means of artificial intelligence / XIII International Scientific and Technical Conference “Computer Science and Information Technologies”, Lviv, Ukraine, 2018 (unpublished)
6. Rajaraman A., Ullman J. D. Data Mining. http://i.stanford.edu/~ullman/mmds/ch1.pdf. Accessed Mar. 24, 2018.
7. Stemming and lemmatization. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.
8. Sparck J.K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval // Journal of Documentation, 1972, vol 28, pp.11–21.
9. Kowsari K., Heidarysafa M., Brown D., Meimandi J.K., Barnes L.E. RMDL: Random Multimodel Deep Learning for Classification // arXiv.org e-Print archive. rXiv:1805.01890 Freely accessible. 2018.
10. Yarovyi A., Timchenko L., et al. Parallel-hierarchical processing and classification of laser beam profile images based on the GPU-oriented architecture / Proc. SPIE 10445, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017, 104450R. doi: 10.1117/12.2280975
11. Timchenko L., Yarovyi A., et al. The method of parallel-hierarchical transformation for rapid recognition of dynamic images using GPGPU technology / Proc. SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016, 1003155. doi: 10.1117/12.2249352