2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)

Fake News Detection Using Naive Bayes Classifier

Mykhailo Granik, Volodymyr Mesyura
Computer Science Department
Vinnytsia National Technical University
Vinnytsia, Ukraine
Fcdkbear@gmail.com, vimes2009@yandex.ru

***Abstract* — This paper shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. We achieved classification accuracy of approximately 74% on the test set which is a decent result considering the relative simplicity of the model. This results may be improved in several ways, that are described in the article as well. Received results suggest, that fake news detection problem can be addressed with artificial intelligence methods.**

Keywords—fake news; naive Bayes classifier; artificial intelligence

I. INTRODUCTION

Internet and social media made the access to the news information much easier and comfortable. Often Internet users can follow the events of their interest in online mode, and spread of the mobile devices makes this process even easier.

But with great possibilities come great challenges. Mass-media have a huge influence on the society, and as it often happens, there is someone who wants to take advantage of this fact. Sometimes to achieve some goals mass-media may manipulate the information in different ways. This leads to producing of the news articles that are not completely true or even completely false. 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 websited may be found in Ukraine, United States of America, Germany, China and lots of other countries [1]. Thus, fake news is a global issue as well as a global 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.

There are several influential articles about automatic deception detection. In [3] the authors provide a general overview of the available techniques for the matter. In [4] the authors describe their method for fake news detection based on the feedback for the specific news in the microblogs. In [5] the authors actually develop two systems for deception detection based on support vector machines and naive Bayes classifier (this method is used in the system described in this paper as well) respectively. They collect the data by means of asking people to directly provide true or false information on several topics – abortion, death penalty and friendship. The accuracy of the detection achieved by the system is around 70%.

This article describes a simple fake news detection method based on one of the artificial intelligence algorithms – naive Bayes classifier. The goal of the research is to examine how this particular method works for this particular problem given a manually labeled news dataset and to support (or not) the idea of using artificial intelligence for fake news detection. The difference between these article and articles on the similar topics is that in this paper naive Bayes classifier was specifically used for fake news detection; also, the developed system was tested on a relatively new data set, which gave an opportunity to evaluate its performance on a recent data.

II. SPAM MESSAGES. SIMILARITY BETWEEN SPAM MESSAGES AND FAKE NEWS ARTICLES

Electronic spamming is the use of electronic messaging systems to send an unsolicited message (spam), especially advertising, as well as sending messages repeatedly on the same site [6].

Spam messages and fake news articles have a lot of common properties:

- They often have a lot of grammatical mistakes.
- They are often emotionally colored.
- They often try to affect reader's opinion on some topics in manipulative way.
- Their content is often not true (this property holds for the most of spam messages and for all of the fake news by definition).
- They often use similar limited set of words. Please note, that this claim is not about the fact, that spam messages and fake news articles use similar set of words. This claim is about the fact, that different spam messages often look like the other spam messages from the syntactic point of view. The same property holds for fake news articles.

So one can see, that fake news articles and spam messages indeed share a lot of important properties. Therefore, it makes sense to use similar approaches for spam filtering and fake news detection.

978-1-5090-3006-4/17/$31.00 ©2017 IEEE

III. Naive Bayes classifier and its usage for spam filtering

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [7].

Naive Bayes classifiers are a popular statistical technique of e-mail filtering. They emerged in the middle of the 90s and were one of the first attempts to tackle spam filtering problem [8].

Naive Bayes typically use bag of words features to identify spam e-mail, an approach commonly used in text classification. Naive Bayes classifiers work by correlating the use of tokens (typically words, or sometimes other constructions, syntactic or not), with spam and non-spam e-mails and then using Bayes theorem to calculate a probability that an email is or is not a spam message [8].

We used very similar approach for fake news detection, and it will be described in more details in the next chapter.

I. Mathematical model of naive Bayes classifier for fake news detection

The main idea is to treat each word of the news article independently.

As were already mentioned, fake news articles often use the same set of words, which may indicate, that the specific article is indeed a fake news article. Of course, it is impossible to claim that the article is a fake news just because of the fact, that some words appear in it, but these words affect the probability of this fact.

The formula for calculating the conditional probability of the fact, that news article is fake given that it contains some specific word looks as following:

$$Pr(F|W) = Pr(W|F)\cdot Pr(F)/(Pr(W|F)\cdot Pr(F)+Pr(W|T)\cdot Pr(T)), \quad (1)$$

where:

Pr(F|W) – conditional probability, that a news article is fake given that word W appears in it;

Pr(W|F) – conditional probability of finding word W in fake news articles;

Pr(F) – overall probability that given news article is fake news article;

Pr(W|T) – conditional probability of finding word W in true news articles;

Pr(T) – overall probability that given news article is true news article.

This formula is derived from Bayes' theorem.

Consider that probabilities Pr(F|W) are known for each word of the news article. Next step is combining this probabilities to get the probability of the fact, that given news article is fake. The formula for this looks as following:

$$p1 = Pr(F|W1)\cdot\ldots\cdot Pr(F|Wn), \quad (2)$$

$$p2 = (1 - Pr(F|W1))\cdot\ldots\cdot(1 - Pr(F|Wn)), \quad (3)$$

$$p = p1/(p1 + p2), \quad (4)$$

where:

n – total number of words in the news article;

p1 – product of the probabilities that a news article is fake given that it contains a specific word for all of the words in the news article;

p2 – same as p1, but complement probabilities are used instead;

Pr(F|W1), Pr(F|W2) … Pr(F|Wn) – conditional probabilities that a news article is a fake given that words W1, W2, Wn respectively appear in it;

p – the overall probability of the fact that given news article is fake.

This formula is often used for spam filtering. [8]

The last question is how to calculate the conditional probabilities of finding specific word in fake news articles and in true news articles. Consider there is a training set, that contains lots of news articles, labeled as true or fake. Then one can define the probability of finding specific word in fake news article as a ratio of the fake news articles, that contain this word to the total number of fake news articles. The probability of finding specific word in true news articles can be defined similarly.

IV. Overview of the training dataset

Dataset, collected by BuzzFeed News [9], was used for learning and testing the naive Bayes classifier.

The dataset contains information about Facebook posts, each of which represent a news article. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). All nine pages have earned the coveted verified blue checkmark from Facebook, which gives them an additional layer of credibility on the platform. The smallest of these public pages has over 450 thousand of followers, and the largest – over 4.1 million followers [9].

BuzzFeed news employees logged and fact-checked each of the posts, that was published on these pages during seven weekdays. They labeled each of the posts as "mostly true", "mostly false", "mixture of true and false" and "no factual content". They also gathered additional data: Facebook engagement numbers (shares, comments, and reactions) for each post were added from the Facebook API. They also noted whether the post was a link, photo, video, or text. Raters were asked to provide notes and sources to explain their rulings of "mixture of true and false" or "mostly false." They could also indicate whether they were unsure of a given rating, which would trigger a second review of the same post in order to ensure consistency. Any discrepancies between the two ratings were resolved by a third person. That same person conducted a

final review of all posts that were rated mostly false to ensure they warranted that rating [9].

In the end, BuzzFeed team rated and gathered data on 2282 posts. There were 1145 posts from mainstream pages, 666 from right-wing pages, and 471 from left-wing pages. The difference in the number of posts for each group is a result of them publishing with different frequencies [9].

V. Implementation Details

The relevant implementation details are the following:

- Among the fields, that are present in the dataset, only few of them were used. They are link to the Facebook post with the text of the news article and the label of the text.
- Text of the news articles was retrieved using Facebook API.
- News articles with labels "mixture of true and false" and "no factual content" were not considered. Couple of the articles in the dataset are broken – they do not contain any text at all (or contain "null" as a text). These articles were ignored as well. After such filtering data set with 1771 news articles was obtained.
- The dataset was randomly shuffled, and after that divided into three subsets: training dataset, validation dataset, test dataset. Training dataset was used for training the naive Bayes classifier. Validation dataset was used for tuning some global parameters of the classifier. Test dataset was used to get the unbiased estimation of how well the classifier performs on new data (it is a well known fact, that it is not correct to only have training and test datasets when parameter tuning is performed, because received results on test set will be biased in this case).
- For the unconditional probability of the fact, that any news article is correct all of the values from interval [0.2; 0.75] with step 0.01 were considered. For the true probability threshold all of the values from interval [0.5; 0.9] with the same step were considered. The best results on the validation dataset were received with the unconditional probability of the fact, that any news article is correct being equal to 0.59 and the true probability threshold being equal to 0.8.
- The global parameters, that were tuned, are the unconditional probability of the fact, that any news article is correct and the true probability threshold. The true probability threshold is such a value, that every article with probability to be true news article bigger than the threshold would be considered by the classifier as a true news article, and all other articles – as a false news articles.
- Consider the classification procedure of the naive Bayes classifier. When iterating through the words of the news article that is being classified, a corner case is possible: some specific word might not be present in the training dataset at all. For all such words it was decided to define the probability of the news article being fake given that it contains this word as 0.5. Equation (4) won't be affected in such case: indeed, both nominator and denominator get multiplied by 0.5. Basically, current implementation just ignores such words.
- If all of the words in the news article are unknown to the classifier (never occurred in the training dataset), the classifier reports, that it can not classify given news article.
- If some word occurred in the news article several times, it contributed to the total probability of the fact, that a news article is fake exactly the same number of times.
- Equation(4) is computationally unstable if calculated directly. This is caused by the fact, that lots of probabilities get multiplied, and the result of such multiplication becomes close to zero really fast. Most of programming languages do not provide the needed degree of precision, and that's why they interpret the result of multiplication as exactly zero. Let p be the probability of the fact, that given news article is fake. One can calculate the value $1/p - 1$ instead, and after that receive the value of p quite easily. The following equation holds:

$$1/p - 1 = p2 / p1 \quad (5),$$

where p, p1, p2 are the same as in (2), (3) and (4). p1 and p2 can be calculated in more stable way using logarithms and exponentiation.

Figure 1 shows the generalized scheme of the used algorithm.

VI. Received Results

The results, that were received, are shown in the Table 1.

The classification accuracy for true news articles and false news articles is roughly the same, but classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it are fake news.

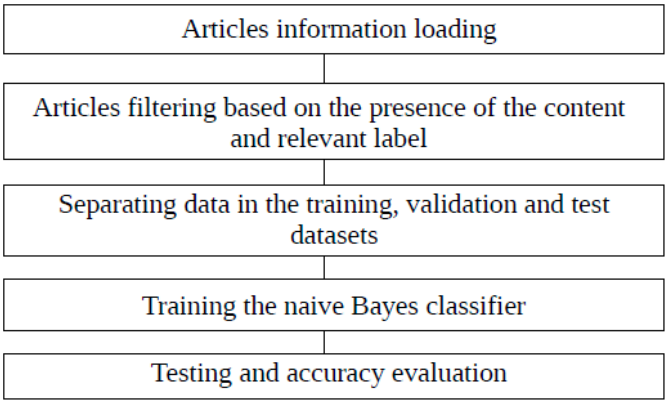

Fig. 1. Generalized scheme of the algorithm

Let's consider the result as positive, when the classifier classifies the news article as fake. Then:

- The number of true positive examples is the number of news articles, correctly classified as fake;
- The number of false positive examples is the number of news articles incorrectly classified as fake;
- The number of true negative examples is the number of news articles, correctly classified as true;
- The number of false negative examples is the number of news articles incorrectly classified as true;

The precision of a classifier is calculated as follows:

$$\text{Precision} = tp / (tp + fp) \quad (6),$$

where:

tp – number of true positive examples;

fp – number of false positive examples.

The recall of a classifier is calculated as follows:

$$Recall = tp / (tp + fn), \quad (7)$$

where fn is a number of false negative examples.

The precision for the given classifier equals to 0.71; recall, on the other hand equals to 0.13. Such a low value of the recall, once again, is caused by the skewness of the data in the test dataset. We believe that precision is the most important characteristic of the given classifier.

TABLE I. RECEIVED RESULTS

News article type	Total number of news in test dataset	Number of correctly classified news	Classification accuracy
True	881	666	75.59%
Fake	46	33	71.73%
Total	927	699	75.40%

II. WAYS TO IMPROVE CLASSIFIER

There are some ways that should improve the performance of classifier. They are as following:

- Get more data and use it for training. In machine learning problems it is often the case when getting more data significantly improves the performance of a learning algorithm. The dataset, that was described in this article contains only around 2000 articles. This number is really small, and we believe that a dataset with couple of millions of news articles would be of a great help for the learning process. Unfortunately, such a dataset is not freely available right now.

- Use the dataset with much greater length of the news articles. The news articles, that were presented in the current dataset, usually were not that long, because they often were just a preview to a longer news article, available on the website, different from Facebook. Training a classifier on a dataset with larger news articles should improve its performance signifacntly.

- Remove stop words from the news articles. Stop words are the words, that are common to all types of texts (such as articles in English). This words are so common, that they don't really affect the correctness of the information in the news article, so it makes sense to get rid of them [10].

- Use stemming. In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form – generally a written word form [11]. Such technique helps to treat similar words (like "write" and "writing") as the same words and may improve classifier's performance as well.

- Treat rare words separately. The words, that are rare to the dataset (occurred less than N times for some constant N) may affect the calculations greatly, but in fact they do not provide much information. It could make sense to just ignore such words.

- Use group of words instead of separate words for calculating probabilities. This will help to use more meaningful syntax constructions for Bayes classifier, but it also requires larger dataset with longer news articles.

Lot's of this improvements are used in spam filtering as well [8].

The exact effect of this improvements should be a subject of a further research, but all of them look promising in our opinion.

Of course, classification accuracy probably would be significantly improved by means of using more complex model. It is worth noting, that even with given dataset, only part of the information was used. Such important information, as number of likes or shares may correlate with news correctness as well.

VII. CONCLUSION

The research showed, that even quite simple artificial intelligence algorithm (such as naive Bayes classifier) may show a good result on such an important problem as fake news classification. Therefore the results of this research suggest even more, that artificial intelligence techniques may be successfully used to tackle this important problem.

REFERENCE

[1] Fake news websites. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017.

[2] Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news. [Online]. Available: https://www.wired.com/2016/12/bittersweet-sweepstakes-build-ai-destroys-fake-news/

[3] Conroy, N., Rubin, V. and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.

[4] Markines, B., Cattuto, C., & Menczer, F. (2009, April). Social spam detection. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 41-48)

[5] Rada Mihalcea , Carlo Strapparava, The lie detector: explorations in the automatic recognition of deceptive language, Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, August 04-04, 2009, Suntec, Singapore

[6] Spamming. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Spamming. Accessed Feb. 6, 2017.

[7] Naive Bayes classifier. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier. Accessed Feb. 6, 2017.

[8] Naive Bayes spam filtering. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering. Accessed Feb. 6, 2017.

[9] Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, Jeremy Singer-Vine. (2016, Oct. 20). Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. [Online]. Available: https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis?utm_term=.twM44ywz1B#.cxEnnGWD6g

[10] Stop words. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Stop_words. Accessed Feb. 6, 2017.

[11] Stemming. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Stemming. Accessed Feb. 6, 2017.