DOI: https://doi.org/10.33216/1998-7927-2019-253-5-31-36

UDC: 004.048

MODEL-ORIENTED FAKE NEWS DETECTION ON SOCIAL MEDIA

Davidenko M.O., Biloborodova T.O.

МОДЕЛЬ-ОРІЄНТОВАНИЙ ПІДХІД ДО ВИЗНАЧЕННЯ ФЕЙКОВИХ НОВИН В СОЦІАЛЬНИХ МЕРЕЖАХ

Давіденко М.О., Білобородова Т.О.

Nowadays, fake news (FN) have actively penetrated throughout the social media reducing our ability to critical assess and proceed the information. Most of existing approaches to handle with FN require a labeled FN training datasets but in some cases these datasets are unavailable. In this paper, we present a model-oriented approach for FN detection and feature extraction. The unsupervised technique for FN identification without the training data is designed and developed. It includes four main steps, namely data preprocessing, text feature extraction, vectorization, and clustering using k-means algorithm. The results of the last step was evaluated through several parameters: homogeneity, completeness, V-measure, Adjusted Rand index and Silhouette coefficient.

Keywords*: FN detection, text mining, model-oriented approach, clustering, word2vec*

1. Introduction

The researches of Stanford University defined FN like the news that are intentionally and verifiably false and can mislead [1].

After the presidential election with Donald Trump and Hillary Clinton at 2016 in the USA, the actual topic was the “Fake News” [2]. Some political pundits claim that FN affected the election. Fake news posts have used social media to disseminate during the internet.

Nowadays, big social networking services companies are developed solutions for FN recognizing. For example, Facebook allows users scoring the news that is possibly suspicious [3]. Recently, a new online service “Google News Initiative” that is proposed by Google to fighting FN [4].

There are many researches provided in this area. The authors [5] investigated and characterized FN and data related to it as follows (Fig.1).

The research approaches for FN detection can be defined in the following categories [6]: data-oriented, feature-oriented, model-oriented, application-oriented.

According to our goal, we detailed investigation and analysis of the model-oriented FN research.

Most approaches include extracting various features and follow using these features into supervised classification models [7, 8]. The conjunction of several weak classifiers into ensemble methods is more successful than any single classification model alone. The ensemble methods have been widely applied to FN detection and have a more accurate result [9].

However, the accuracy of FN detection, is still challenging, due to the dynamic nature of social media, and the complexity and diversity of online text data. Also, in the absence of high-quality training data is a problem for the creation of detection models.

One of the major challenges for FN detection is the fact that each feature, such as source trustworthiness, style of news text, or social response, has some restriction to directly predict FN singly. The process of obtaining a reliable FN dataset is difficult for the following reasons [10]:

- the real-world online dataset is usually big, incomplete, unstructured, unlabeled, and noisy;
- everyday a large amount of false information with diverse intentions and different linguistic characteristics is created via social media.

Also, most existing approaches require a labeled FN train dataset to train a model. It is important to consider scenarios where limited or no labeled FN items are available in which semi-supervised or unsupervised models can be applied. The models created by supervised learning be more accurate given a quality training dataset, unsupervised models can be more useful on wide availability of unlabeled data.

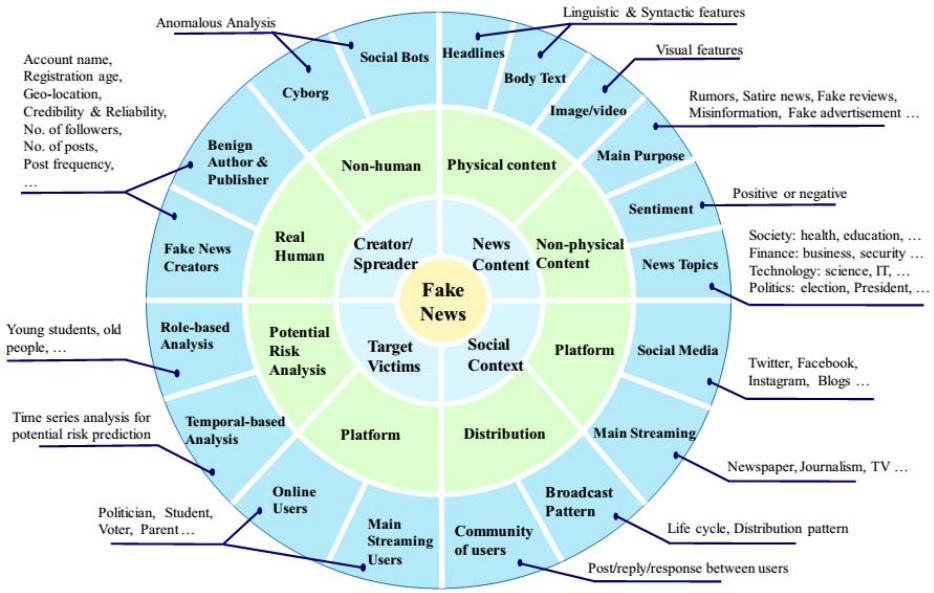

Fig. 1. Characteristic of FN and data related to it [5]

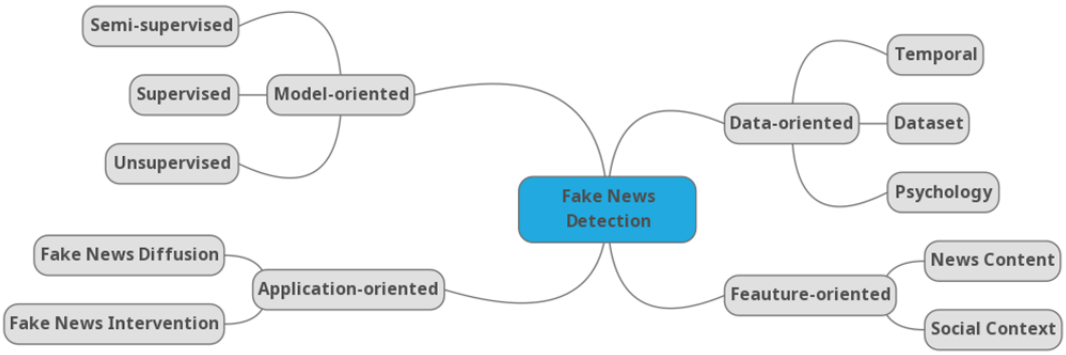

Fig. 2. The categories of research approaches for FN detection

It is necessary to design an approach which is able to identify FN even without the training samples. It is important to design effective, automatic and applicable approaches for high-quality online FN detection.

The goal is effectiveness model-oriented approaches for FN detection. The objectives can be defined as follow:

- study and analyzed related work for FN detection and feature extraction;
- designed and development of unsupervised technique for FN detection consisting of data preprocessing, text feature extraction, converting words to vectors and clustering using k-means algorithm;
- clustering evaluation.

The outline of the paper is as follows. In Section 2, we describe the major definitions of model-oriented approaches for FN detection. The method of feature extraction is considered in Section 3. The model-oriented method such as clustering is provided and experiment process is presented in Section 4. The clustering evaluations are present in Section 5. Finally, we conclude our research.

2. Definitions of model-oriented approaches

According to the model-oriented approach for FN detection can be used unsupervised and supervised machine learning technique. We study classification as supervised learning, and perform clustering as unsupervised learning.

The clustering in the context of FN detection can be defined as follow [11]. Given a corpus of F N $J = \{j_1, j_2, j_3, \ldots, j_n\}$ with size of n where each document $\overrightarrow{n_i}$ is a vector of terms in a dictionary, $\sum = \{t_1, t_2, t_3, \ldots, t_T\}$ with size of $T = |\sum|$. The problem is clustering of documents based on their terms into homogeneous classes with respect to FN categories. To this end, we first cluster documents based on appearance positions of each term in an article and its correlations with other terms (Spatial relation extraction) following by designing an automatic ensemble co-clustering to cluster documents according to their positions in different factors among various low-rank decompositions.

Mathematically, the classification for FN detection define as follows [12]. A news item is called fake if its content is verified to be fake and true otherwise. Let $X = \{X_1, X_2, ..., X_n\}$ denote a dataset containing *n* news items. Each new item $j \in [1, n]$ contains *k* resources of data and is denoted as $X_j = \{X_j^1, X_j^2, ..., X_j^k\}$. Additionally, let $Y = \{y_1, y_2, ..., y_n\}$ is a set of class labels related with news samples of dataset *X*. Each class label $y_i \in Y$ from the label set, where *m* is the number of recognized class of news in dataset and $l_j \in L$ is a class meaning: fake or true.

With the aforementioned notations and definitions, the problem of multi-source multi-class FN detection is formally defined as follows. Given the dataset *X* and its corresponding labels *Y*, the goal of classification is to learn the model *M* mapping *X* to *Y*, which predicts the degrees of fakeness for unlabeled news.

3. Feature extraction

The wor2vec model is a popular technique for feature extraction in text mining. It can be describing as follow. The word2vec model takes a large text corpus as input and maps each word to a vector, giving the coordinates of the words in the output. At the first stage, it creates a dictionary through learning on the input text data, and then calculates the vector of words. The vector is based on contextual proximity: words found in the text follow to identical words (therefore having similar meanings) in the vector have close coordinates of word vectors. The parameters and corpus sizes are affected to the model accuracy. Accuracy increases overall if the number of words used increases, and if the number of dimensions increases [13].

The optimization of word2vec model is a classification problem [14] and defineded as follow. The word embedding layer is a matrix of a number of unique words in the corpus and words embedding size. Suppose a textual source contains *x* words [12]. The neural network model applying, the text represents by an input matrix of word embeddings denoted as $W \in R^{x \times e}$ where e is the dimension of the word embedding. More specifically, $w_j \in W$ is a e dimensional vector representing *j*-th word of the text and populates *j*-th row of matrix *W*. In other words, each row of the matrix represents a word in the corpus. Words embedding size is a hyperparameter to be decided and how many features that can be defined to each word. The last stage of the model is a logistic regression in a neural network form.

4. Experiment

4.1. Data Description

The dataset by George McIntire [15] was used. The dataset was prepared in 2017. It consists of data from 5279 articles. The articles came from media organizations such as the New York Times, WSJ, Bloomberg, NPR, and the Guardian and were published in 2015 or 2016. The dataset consists of the headline and text of a news article as input variables and output variable with the two classes: FAKE or REAL.

The FN detection includes follow steps.

4.2 Data preprocessing

The raw data was preprocessed. The bad characters, tokenize and stop words are removed. The raw data is present as follow (see Fig.3).

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

Fig. 3. Data before processing

	Unnamed: 0	title	text	label
0	8476	[smell, hillarys, fear]	[daniel, greenfield, shillman, journalism, fel...	FAKE
1	10294	[watch, exact, moment, paul, ryan, committed, ...	[google, pinterest, digg, linkedin, reddit, st...	FAKE
2	3608	[kerry, go, paris, gesture, sympathy]	[us, secretary, state, john, f, kerry, said, m...	REAL
3	10142	[bernie, supporters, twitter, erupt, anger, dn...	[kaydee, king, kaydeeking, november, lesson, t...	FAKE
4	875	[battle, new, york, primary, matters]	[primary, day, new, york, frontrunners, hillar...	REAL
5	6903	[tehran, usa]	[im, immigrant, grandparents, years, ago, arri...	FAKE
6	7341	[girl, horrified, watches, boyfriend, left, fa...	[share, baylee, luciani, left, screenshot, bay...	FAKE
7	95	[britains, schindler, dies]	[czech, stockbroker, saved, jewish, children, ...	REAL
8	4869	[fact, check, trump, clinton, commanderinchief...	[hillary, clinton, donald, trump, made, inaccu...	REAL
9	2909	[iran, reportedly, makes, new, push, uranium, ...	[iranian, negotiators, reportedly, made, lastd...	REAL

Fig. 4. Data after preprocessing step

Data after preprocessing without bad characters, tokenize and stop words can be present as follow (see Fig.4).

4.3. Train word2vec model

After training the model with the data generated from the example sentence above, we can see that the model can output most of the similar words for each word as an input word.

Take the sentence for example, given a context word "go" we would like the model to generate one of the underlying words (one of the words in [stay, come, get, sit, throw, wait, happen, alone, roll, let] in the follows case.).

model.wv.similar_by_word("go")

```
[
 ('stay', 0.7747250199317932),
 ('come', 0.7660486698150635),
 ('get', 0.751580536365509),
 ('sit', 0.7376989722251892),
 ('throw', 0.7257541418075562),
 ('wait', 0.7247833013534546),
 ('happen', 0.6911652088165283),
 ('alone', 0.6910232305526733),
 ('roll', 0.6893036365509033),
 ('let', 0.6857764720916748)
]
```

On another case, given a context word “politic” we would like the model to generate one of the underlying words (one of the words in [destructive, healthcare-aboveall, karma, potent, distributionthe, interlocking, ecologically, intolerable, divine, oneness]).

model.wv.similar_by_word("politic")

```
[
 ('destructive', 0.8911725282669067),
 ('healthcareaboveall', 0.888471782207489),
 ('karma', 0.8859128952026367),
 ('potent', 0.8848767876625061),
 ('distributionthe', 0.8825998902320862),
 ('interlocking', 0.8820247054100037),
 ('ecologically', 0.8805824518203735),
 ('intolerable', 0.8803092837333679),
 ('divine', 0.8801225423812866),
 ('oneness', 0.8792058229446411)
]
```

The follow steps after word2vec are converting words to vectors and using K-Means algorithm for clustering.

5. Clustering evaluation

There are many parameters are used for clustering evaluation. The follow parameters are used in our experiment: homogeneity, completeness, V-measure, silhouette, Adjusted Rand index, Adjusted Mutual Information and Silhouette coefficient.

Formally, homogeneity h, completeness c, V-measure V are determined using the functions of entropy as follow.

$$h = 1 - \frac{H(C|K)}{H(C)}, c = 1 - \frac{H(K|C)}{H(K)},$$

where K is the result of clustering, C is the true partitioning of the sample into classes. Thus, homogeneity measures how much each cluster consists of objects of the same class, and completeness measures how much objects of the same class belong to the same cluster. These measures are not symmetrical. Both values take values in the range [0,1], and large values correspond to more accurate clustering. These measures are not normalized and depend on the number of clusters.

To take into account homogeneity and completeness, a V-measure is introduced at the same time as their harmonic mean. It is symmetrical and shows how the two clusters are similar to each other.

The V-measure calculated as follow.

$$V = (1 + beta) * h * c / (beta * h + c)$$

The Rand index (RI) expresses the similarity of two different clusters of the sample. In order for this index to give values close to zero for random clustering for any n and the number of clusters, it is necessary to normalize it. This is how the Adjusted Rand Index is defined. Adjusted Rand index (ARI) is a symmetric measure, independent of label values and permutations. This index is a measure of the distance between different sample partitions. ARI takes values in the range [−1,1]. Negative values correspond to "independent" clusters, values close to zero correspond to random partitions, and positive values indicate that the two partitions are similar. The Rand Index and Adjusted Rand index can be defined as follow.

$$RI = \frac{2(a+b)}{n(n-1)}, \; ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]},$$

where n is the number of objects in the set, a is the number of both of objects that have the same labels and are in the same cluster, b is the number of pairs of objects that have different labels and are in different clusters.

Silhouette coefficient is a method of interpretation and validation of data consistency in clusters.

The silhouette value is a measure of how similar an object is to its own cluster is it compare to other clusters. It takes values as follow [−1, 1]. A high value means that the object is well similar to its own cluster and badly similar to other clusters.

The results of evaluations parameters are:

Homogeneity: 1.000
Completeness: 1.000
V-measure: 1.000
Adjusted Rand Index: 1.000
Silhouette Coefficient: 0.326

Conclusion

The most of approaches require a labeled data for accurate FN detection. The models created by supervised classification methods may be more accurate given a high-quality dataset for training. However, unsupervised models don't require a labeled data and can be more practical because unlabeled datasets are more available to obtain. The unsupervised technique to identify FN without the training data is designed and development. It consists of follow steps: data preprocessing, text feature extraction, converting words to vectors and clustering using k-means algorithm.

We provide clustering evaluation through several parameters: homogeneity, completeness, V-measure, silhouette, Adjusted Rand index and Silhouette coefficient. Based on the results of the Silhouette clustering coefficient as the main evaluation method, we can conclude that the clustering model is accurate.

Follow clustering evaluation, the word2vec model showed quality results of vectorizing and searching similar words to build distances between objects for the follow clustering analysis.

The future research is additional settings such as doc2vec method for feature extraction and cross-validation for find better configurations for clustering model accuracy improving.

References

1. Detecting fake news with nlp [Electronic resource]. – Avaible at : https://medium.com/@Genyunus/detecting-fake-news-with-nlp-c893ec31dee8. Accessed: 2019-04-15.
2. How to Build a "Fake News" Classification Model - Open Data Science - Your News Source for AI, Machine Learning & more [Electronic resource]. – Avaible at : https://opendatascience.com/how-to-build-a-fake-news-classification-model/. Accessed: 2019-04-15.
3. News Feed fyi. Addressing hoaxes and fake news [Electronic resource]. – Avaible at : https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/. Accessed: 2019-04-15.
4. Google News Initiative [Electronic resource]. – Avaible at : https://newsinitiative.withgoogle.com/. Accessed: 2019-04-15
5. Zhang, X. and Ghorbani, A.A., 2019. An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management.
6. Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), pp.22-36.
7. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. and Stein, B., 2017. A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638.
8. Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S. and de Alfaro, L., 2017. Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506.
9. Dietterich, T.G., 2000, June. Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.
10. Zhang, X. and Ghorbani, A.A., 2019. An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management.
11. Hosseinimotlagh, S. and Papalexakis, E.E., 2018. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. MIS2, Marina Del Rey, CA, USA.
12. Karimi, Hamid, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. "Multi-source multi-class fake news detection." In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1546-1557. 2018.
13. Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
14. I. Chen Word2vec from Scratch with NumPy. Towardsdatascience.com [Electronic resource]. – Avaible at: https://towardsdatascience.com/word2vec-from-scratch-with-numpy-8786ddd49e72. Accessed: 2019-04-15.
15. Github.com [Electronic resource]. – Avaible at : https://github.com/ricknta/fake-news. Accessed: 2019-04-15.

Давіденко М.О., Білобородова Т.О. Модель-орієнтований підхід до визначення фейкових новин в соціальних мережах

На даний час фальшиві новини (FN) активно проникають в соціальні мережі, знижуючи нашу здатність критично оцінювати і обробляти інформацію. Більшість існуючих підходів для роботи з FN вимагають маркованих навчальних наборів даних FN, але в деяких випадках ці набори даних недоступні. У цій статті ми представляємо модельно-орієнтований підхід для виявлення FN і виділення ознак. Неконтрольована методика ідентифікації FN без навчальних даних спроектована і розроблена. Вона включає в себе чотири основні етапи: попередню обробку даних, вилучення текстових ознак, векторизацію і кластеризацію з використанням алгоритму k-середніх. Результати останнього етапу оцінювалися за кількома параметрами: однорідність, повнота, V-мера, скоригований індекс Ренді і коефіцієнт силуету.

Ключові слова: *визначення фейк новин, інтелектуальний аналіз тексту, модельно-орієнтований підхід, кластеризація, word2vec.*

Давыденко Н.А., Белобородова Т.А. Модель-ориентированный подход для определения фейковых новостей в социальных сетях

В настоящее время фальшивые новости (FN) активно проникают в социальные сети, снижая нашу способность критически оценивать и обрабатывать информацию. Большинство существующих подходов для работы с FN требуют маркированных обучающих наборов данных FN, но в некоторых случаях эти наборы данных недоступны. В этой статье мы представляем мо-

дельно-ориентированный подход для обнаружения FN и выделения признаков. Неконтролируемая методика идентификации FN без обучающих данных спроектирована и разработана. Она включает в себя четыре основных этапа: предварительную обработку данных, извлечение текстовых признаков, векторизацию и кластеризацию с использованием алгоритма k-средних. Результаты последнего этапа оценивались по нескольким параметрам: однородность, полнота, V-мера, скорректированный индекс Ренда и коэффициент силуэта.

Ключевые слова: *обнаружение фейк новостей, интеллектуальный анализ текста, модельно-ориентированный подход, кластеризация, word2vec.*

Давіденко Микита Олександрович – магістр кафедри комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: nickita.davidenko@gmail.com

Білобородова Тетяна Олександрівна – к.т.н., ст.викладач кафедри комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: beloborodova.t@gmail.com

Стаття подана 6.08.2019.