The Comparison of Local and Global Early Fake News Detection Methods
Project Proposal for NLP Course, Winter 2022

Hubert Ruczyński
WUT
01151402@pw.edu.pl

Bartosz Siński
WUT
01151411@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Maciej Pawlikowski
WUT
01151389@pw.edu.pl

Adrian Stańdo
WUT
01151435@pw.edu.pl

Abstract

In this report, we describe our initial research toward solving the problem of identifying trending rumors on social media platforms. We explore already existing solutions for topic detection, fake news detection, and explainability in the field of NLP. We aim to design a novel rumor detection framework, where the classifiers are trained locally for the most important topics. Such an approach will be topic-aware, and less expensive computationally.

1 Introduction

Nowadays most people learn about the world around them from the resources on the internet. Therefore not surprisingly it is used by the majority of the current news media companies. Unfortunately sometimes among the information provided by several reliable news outlets, there are a few misleading articles in which authors want to fool the reader. Their intentions could be to just grasp the attention of the user or in some cases to manipulate the user for personal gains. This misleading information is called fake news. Detecting such articles and differing them from real ones is a very important initiative. Creating a solution that would effectively recognize fake content could prevent the spread of disinformation and prevent people from harm.

As the solution, our team proposes a novel technique of rumor recognition that finds suspicious content at the early stage of its propagation. In order to distinguish emerging topics among news content we will use a state-of-the-art (SOTA) topics detection method. Furthermore, we will use unsupervised machine learning methods to find topics around which a lot of fake content is created. By combining these two methods our goal is to discover an innovative approach to identifying dubious information before it spreads too extensively. On top of that, we will implement a classifier that will use the information obtained by the cluster analysis. A strong focus will be employed on the explainability of our solution. From both clustering stage and classification we aim to find new insight regarding fake news detection by using various eXplainable AI (XAI) methods.

2 Related Works

2.1 Topic Detection

The first work regarding topic detection Leo et al. (2023) focuses on finding the clusters of tweets, describing similar discussion areas. This task is extremely important, as Twitter is currently the biggest platform enabling free, and uncensored thoughts exchange. We can clearly underline two major contributions of this work: the introduction of a stable clustering, and semantical enhancement of short messages (tweets). The first one tackles a major issue of topic detection, which is a machine learning (ML) task with fairly unstable results, especially because of issues with selecting a proper number of clusters, which results in chaotic transfers of observations from one cluster to another. An answer to this problem is the usage of Non-Negative Matrix Factorization (NMF) with consensus clustering. The idea behind consensus clustering is that by repeating the clustering operation many times with varying NMF regularization parameters, the words that will stay most of the time in the same cluster are likely to be the correct cluster members. The authors additionally point out various important attributions of tweets, connected to their length. As their maximum length is 280 characters (before - 140), we can assume that a single tweet can carry only one topic, which is a very small assumption. However, it also indicates some drawbacks, as a singular tweet corpus is rather small, and it is hard to carry its true meaning. The paper introduces a semantical enrichment strategy, where we select the most important words, and with the usage of embeddings add similar variations of them, so a singular tweet can carry more information.

Another important work Lossio-Ventura et al. (2019) in this area, compares various LDA approaches, and suggests the data preprocessing options applicable to topic detection for tweets task. The paper presents how to efficiently use Calinski-Harabasz index Caliński and JA (1974), and Silhouette Coefficient Rousseeuw (1987) for clustering evaluation, and shows us, that for this kind of data, GibbsLDA Wei and Croft (2006), and Online Twitter LDA Lau et al. (2012), prove to be better than their counterparts.

2.2 Fake News Detection

There is a multitude of works dedicated to fake news detection describing a lot of ways to approach this subject. For example in Kasra Majbouri Yazdi (2020) authors focus on feature selection based on computing similarity between primary features in the fake news dataset, clustering obtained features using K-means Guo et al. (2004), and selection of final attributes of all clusters. The paper describes in detail how all algorithms used are calculated and presents the value of the proposed feature selection method using it combined with SVM Evgeniou and Pontil (2001) to achieve very good results when it comes to fake news detection.

Another approach proposed in Tian and Baskiyar (2021) not only allowed authors to achieve high accuracy by utilizing Genetic and evolutionary Feature Selection and KNN in the fake news detection but also tested the quantum version of k- nearest neighbors. This research went in depth when it comes to testing the above-mentioned methods on the BuzzFace Williams and Santia (2018) dataset which consists of 2282 news articles and posts about the 2016 election from Facebook, which was divided into several categories: mostly fake, fake, mostly true, true, and mixed true and fake.

2.3 Explainability

Most of the SOTA models developed for the fake news detection task aim to have the greatest performance and accuracy on selected datasets. Lately however as shown by A.B. et al. (2023) new techniques and methods have been created that focus on gaining better insight into the model decision-making process. Authors argue that explaining model predictions is the key gate-away to achieving better results. The authors present 11 SOTA explainable fake news detection methods of which 7 are attention-based approaches.

One example Kurasinski and Mihailescu (2020) of the attention-based method visualizes attention weights as the color-coded text to show the impact of the particular words on the prediction. For the detection task authors use two deep learning models. First is the BiDir-LSTM-CNN which is the architecture that combines convolutional neural networks and bidirectional recurrent neural networks. The second one is bidirectional encoder representation from transformers (BERT Devlin et al. (2019)). Used color-coded visualization to show how models distribute attention differently. Another interesting finding was that both models showed a strong correlation between click-bait content such as *Check it out!, MOST IMPORTANT* and fake news. Models were trained on the "Fake News Corpus" Pathak and Srihari (2019) which is the data set we are using in our solution. For both models preprocessing methods: *summarization, stemming* and *lemmatization* worsened the results.

3 Dataset Description

Online news and posts can be collected from a variety of sources via dedicated APIs or by scrapping. Nonetheless, manual annotation is a challenging task requiring annotators with domain expertise. For these reasons, we will make use of open-source data available on the Internet. As for now, not many datasets, regarding fake news detection, are available, as well as there is not one commonly used benchmark dataset. In this section, we will present a few data sources with short descriptions and a list of shortcomings.

1. *BuzzFeedNews* - data of news published on Facebook during a week before the US election in 2016. Every post was annotated by 5 people, however, the dataset contains only links to Facebook posts containing links (contents of articles are not available).

2. *CREDBANK* - the dataset described in Mitra and Gilbert (2021) contains 60 million tweets that cover 96 days in 2015. Access to the data

is restricted - they have to be downloaded from the AWS cloud for a small fee. Moreover, labels are not provided for the tweets - only events were identified and annotated by 30 different people.

3. *FakeNewsCorpus* - this dataset includes around 9 million news articles (around 30 GB of data). It contains, among other things, the text of the articles, their title, and source. It also provides labels for 9 different types of misinformation with the addition of class *Credible*. It was created by scrapping text from more than 1000 different Internet domains. Each article has been attributed the same label as the label associated with its domain.

There are many other data sources, however, they are smaller and, hence, it may be difficult to create a reliable model using them. All things considered, we made a decision to use the last dataset - *FakeNewsCorpus* - as it is the most extensive one and contains long enough article content.

4 Methods Review

Fake news detection is a complex task that can be solved in a multitude of approaches, but the first approach is data encoding, as machine learning algorithms work on numbers not on letters. Techniques that were proven to be successful in fake news detection were **CountVectorizer**, **bag-of-words** and **TF-IDF** (Term Frequency - Inverse Document Frequency), which assigns each word score equal to number of occurrences of given word in document divided by its length multiplied by 1 + log(percentage of documents containing this word).

The most complex part comes after initial encoding because state-of-the-art (SOTA) approaches mostly devise innovative preprocessing techniques. Authors of Leo et al. (2023) tackle a similar problem to fake news detection- topic detection. To solve this task they propose using **Non-Negative Matrix Factorisation** and the consensus clustering. They begin by approximating the bag-of-words matrix using two lower-rank matrices V and H which represent respectively the document-topic matrix and topic-words matrix. They are calculated to minimize Frobenius norm: $||W - V \cdot H||_l^2$ under the condition that both V and H contain only positive values. Finally from the recovered topic-word matrix H authors select a subset of the most important words in each topic(row). This approach based on NMF factoring is very fast and robust for short text such as tweets and other social media messages. Unfortunately, it is not immune to sparsity problems, which can be solved by changing the document-term matrix into the term-term matrix. It is also challenging to find an optimal number of topics, as there are no clear indications of their optimal number. One approach is to check the stability of the most important words selected while changing the regularization parameter. This is a similar idea to **consensus clustering** which assesses stability by repeating the clustering operation many times with varying NMF regularization parameters and analyzing the words that will stay most of the time in the same cluster. They are likely to be the correct cluster members.

In Lossio-Ventura et al. (2019) authors present a different approach that uses the most popular short text methods. Mainly: **Latent Semantic Indexing** (transforming a term-document matrix using singular value decomposition), **Latent Dirichlet Allocation** (assumes that each document is a mixture of topics, and each topic is a mixture of words, allowing it to discover the underlying themes and structures), LDA with Gibbs Sampling (GibbsLDA), Online LDA, **Biterm** (capture short-range word associations and extract meaningful topics) and Online Twitter LDA.

Authors of Tian and Baskiyar (2021) propose a solution that uses **GEFeS**(Genetic and Evolutionary Feature Selection), which assigns a numeric value for each feature from 0 to 1, based on which a feature is selected if it exceeds a given threshold. This value is calculated using a Genetic algorithm to evolve the Feature Mask to maximize the accuracy of the KNN model for different k-values separately. As quantum computers are prophecized to be faster than standard computers authors also checked the quantum version of the KNN algorithm, which even though should work just like the standard version of the algorithm returned worse results.

An interesting approach was proposed by authors of Kasra Majbouri Yazdi (2020), who decided to compute similarity among primary features of their dataset, used k-means clustering on them, and from each cluster chose one representative feature that contained the most information about the target variable. This allowed them to greatly reduce the number of features in their dataset, which made model training faster and less biased, which can be seen in the results they achieved with the help of SVM.

Not all fake news detection approaches use intricate preprocessing and get their results by using less-known machine learning algorithms. For example, authors of Ahmed et al. (2022) get their best results with the help of a **Passive-aggressive classifier**. For this algorithm, data comes in sequential order, and it makes predictions step by step not in batches, which makes it great for work with text. It gets its name because it doesn't change the model if the prediction is correct (passive) and does so only when the model makes an error (aggressive).

It is a constant struggle to understand the inner workings of machine learning models, especially ones for NLP as models easiest to explain aren't cutting-edge when it comes to fake news detection. Authors of A.B. et al. (2023) focus on this problem and create a survey of XAI for fake news detection. They discuss all the methods from different aspects: utilizing visual characteristics in explainable fake news detection problems, analysis of existing approaches based on explainability types, Examining current models' explainability from the viewpoint of the different explained, taking the evaluation metrics used in current fake news detection models into account. Some popular techniques include **intrinsic interpretability** (models of simple structure), **post-hoc interpretability** (interpreting model after training), **feature importance**, **permutation feature importance**, **local approximation-based**explanation(approximation of black-box model using white-box model), and **attention-based explainability**. Effects of the last method are shown in greater detail in Kurasinski and Mihailescu (2020) where authors train two NLP models: BiDir-LSTM-CNN and BERT and visualize how much attention a model puts on a given word on practical example using color-coded maps.

5 Proposed Solution

This section will provide detailed explanation and intended plan for the proposed solution.

1. We will perform exploratory data analysis accompanied by data preprocessing. The data set that we are using consists of nine million texts labeled therefore thorough data preparation is crucial for further analytical work.

2. Next, we will employ SOTA topic detection tools, based embeddings, noun chunks, n-grams, named entity recognition, and more. It will allow us to perform a detailed analysis of text from articles and determine a distinctive topic.

3. Evaluation of the clustering quality with SOTA approaches mentioned in the Literature Review. By comparing to cutting-edge techniques we can validate the correctness and efficiency of our solutions.

4. Implementation of SOTA solutions for fake news task detection mentioned in the Literature Review. Used methods will consist of different deep learning architectures (ex. BERT), machine learning models (ex. KNN) and NLP techniques.

5. Training and evaluation of chosen methods on the whole *FakeNewsCorpus* data set and on data within the top largest clusters obtained by previous analysis. Comparison of the results of global and local approaches in terms of performance.

6. Explanations using eXplainable AI (XAI) methods to discover the most important indicators and gain insight into the prediction of both global and local methods. Furthermore, we will check the contents of the obtained clusters.

References

A.B., A., Kumar, S. M., and Chacko, A. M. (2023). A systematic survey on explainable ai applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122:106087.

Ahmed, S., Hinkelmann, K., and Corradini, F. (2022). Development of fake news model using machine learning through natural language processing.

Caliński, T. and JA, H. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Evgeniou, T. and Pontil, M. (2001). Support vector machines: Theory and applications. volume 2049, pages 249–257.

Guo, G., Wang, H., Bell, D., and Bi, Y. (2004). Knn model-based approach in classification.

Kasra Majbouri Yazdi, Adel Majbouri Yazdi, S. K. J. H. W. Z. S. S. (2020). Improving fake news detection using k-means and support vector machine approaches.

Kurasinski, L. and Mihailescu, R.-C. (2020). Towards machine learning explainability in text classification for fake news detection. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 775–781.

Lau, J. H., Collier, N., and Baldwin, T. (2012). On-line trend analysis with topic models: #twitter trends detection topic model online.

Leo, V. D., Puliga, M., Bardazzi, M., Capriotti, F., Filetti, A., and Chessa, A. (2023). Topic detection with recursive consensus clustering and semantic enrichment. *Palgrave Communications*, 10(1):1–10.

Lossio-Ventura, J. A., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., and Bian, J. (2019). Clustering and topic modeling over tweets: A comparison over a health dataset. *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, 2019:1544–1547.

Mitra, T. and Gilbert, E. (2021). Credbank: A large-scale social media corpus with associated credibility annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):258–267.

Pathak, A. and Srihari, R. (2019). BREAKING! presenting fake news corpus for automated fact checking. In Alva-Manchego, F., Choi, E., and Khashabi, D., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 357–362, Florence, Italy. Association for Computational Linguistics.

Rousseeuw, P. (1987). Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65.

Tian, Z. and Baskiyar, S. (2021). Fake news detection using machine learning with feature selection. pages 1–6.

Wei, X. and Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 178–185, New York, NY, USA. Association for Computing Machinery.

Williams, J. and Santia, G. (2018). Buzzface: A news veracity dataset withfacebook user commentary and egos.

A Workload

In the table 1 we present an estimated workload for the project including both future milestones. We will divide the work equally for everyone.

Task	Starting Week	Estimated Workload (hours)
EDA	Week 1	8
Data Preparation	Week 1	16
Initial Clustering and Modelling	Week 2	16
Proper Clustering	Week 3	16
Proper Modelling	Week 3	16
Evaluation	Week 3	4
XAI	Week 4	8
Formalities (paper, poster, etc)	Week 5	16

Table 1: Workload Table