Skoltech

Skolkovo Institute of Science and Technology

Fake News Detection using Tensor-based Approach

by Rosnet Team

Albert Sayapin

Fakhriddin Tojiboev

Farid Davletshin

Problem Statement:

Goal:

To identify if particular news is fake or real using features given by tensor/matrix decompositions

Motivation:

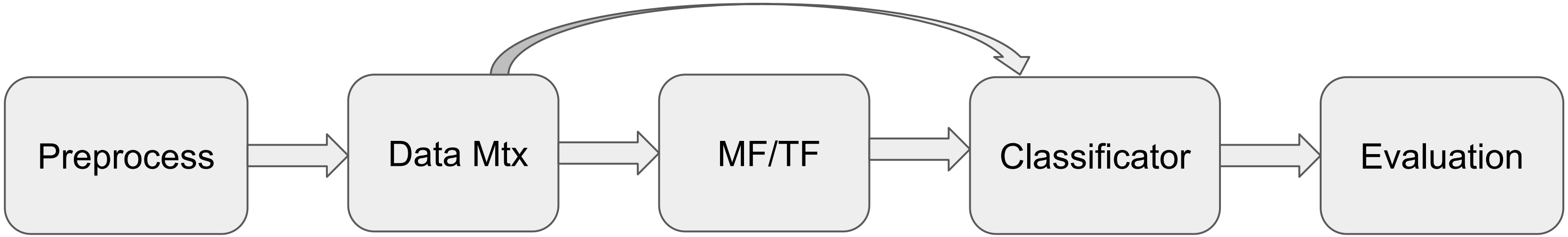
- We have lots of information resources
- Hence, everyday we get different news(especially in the COVID era)
- Some of them are **Fake**
- They were created to manipulate **YOU**
- We are here for the sake of salvation.

Methodology:

Hypothesis:

Latent information based on the social context like user and news article engagement is considered to improve the detection of fake news.

Pipeline:



Methodology:

Data processing:

1. N -> n-gram count mtx: eliminate punct. and numbers, use NLTK WordLemmatizer, use Sklearn CountVectorizer (182 * 3000)
2. U -> news-user mtx: U(i, j) = # j-th user shared i-th article (182 * 15257)
3. D -> user-community mtx: D(i, j) = 1 if i-th user in a j-th community (15257 * 81)

Metrics:

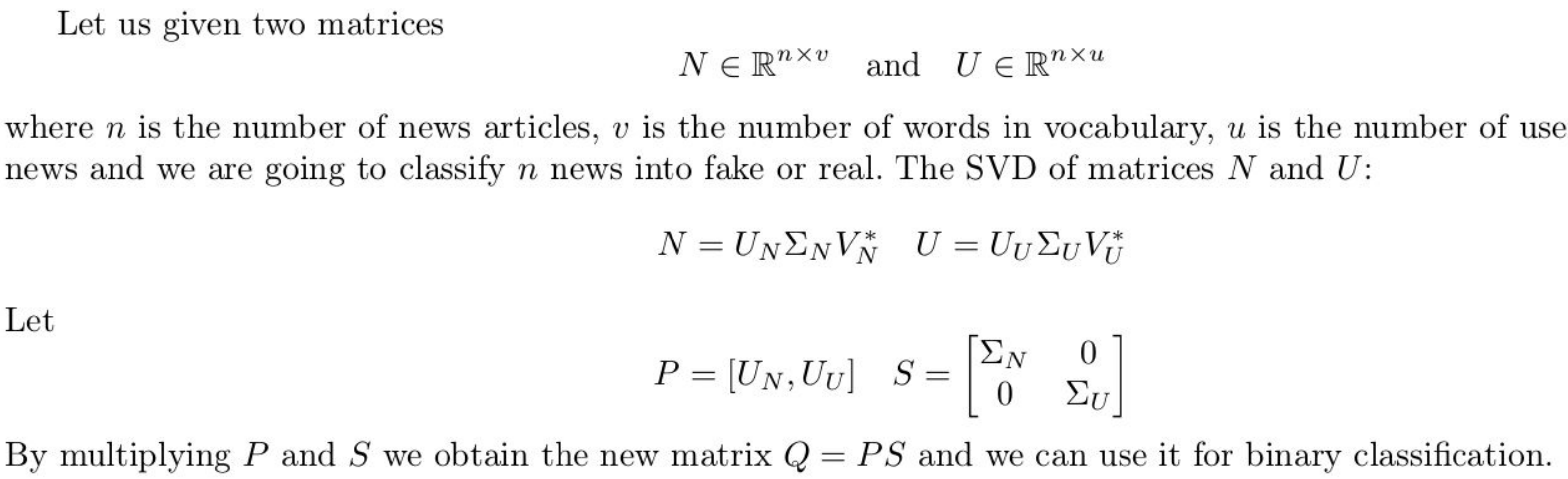
We used these metrics to evaluate the models on a test set:

1. Precision(P)
2. Recall(R)
3. F1 Score(F1)
4. Accuracy(A)

DeepFake model

Methods:

SVD based technique



Let us given two matrices

$$N \in \mathbb{R}^{n \times v} \quad \text{and} \quad U \in \mathbb{R}^{n \times u}$$

where n is the number of news articles, v is the number of words in vocabulary, u is the number of users that shared news and we are going to classify n news into fake or real. The SVD of matrices N and U:

$$N = U_N \Sigma_N V_N^* \quad U = U_U \Sigma_U V_U^*$$

Let

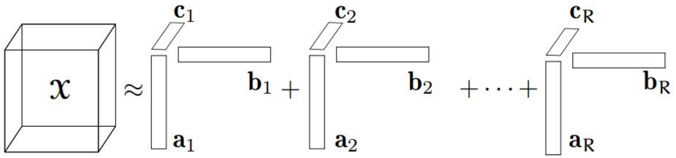
$$P = [U_N, U_U] \quad S = \begin{bmatrix} \Sigma_N & 0 \\ 0 & \Sigma_U \end{bmatrix}$$

By multiplying P and S we obtain the new matrix $Q = PS$ and we can use it for binary classification.

Spearman’s Hypothesis

Methods:

CP Decomposition:



$$\mathfrak{L}(A, B, C, \lambda) = \frac{1}{2}\|\mathcal{X} - \mathcal{M}\|_F^2 + \frac{\lambda}{2}\left(\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2\right) \to \min_{\mathcal{M}}$$

- X is a *news-user-community* tensor s.t. X(i, j, k) = U(i, j)*D(j, k) -> Sparse in COO
- This optimization problem can be rewritten as *3 Least Squares problems*
- Solve them for A, B, C factors sequentially using *python* + *numpy* + *numba*
- Do it several epochs

$$(C^\top C * B^\top B + \lambda I)a_i = (C \odot B)^\top x_i$$

$$i = 1, m$$

U - News-User mtx
D - User-Community mtx

Tensor Decompositions by Kolda

CP and NMF results

Skoltech

Skolkovo Institute of Science and Technology

BuzzFeed: (182 records)

Model	Precision	Recall	F1	Accuracy
Random	0.613	0.594	0.603	0.59
XGB + N1	0.806	0.714	0.757	0.737
XGB + N	0.742	0.741	0.741	0.737
XGB + A	0.838	0.742	0.787	0.77
XGB + A + N1	0.87	0.729	0.794	0.77
XGB + A + N	0.806	0.757	0.781	0.77
DNN + A + N	**0.838710**	**0.896**	**0.866**	**0.868**

- XGB -> XGBoost
- N1 -> NMF first factor mtx
- N -> n-gram count mtx
- A -> 1-mode factor of news-user-community tensor
- DNN -> Deep Neural Network

NMF Cichocki

SVD based technique results:

Skolkovo Institute of Science and Technology

BuzzFeed: (182 records)

Model	P	R	F1	A
XGBoost	0.78	0.78	0.78	0.78
Log reg	0.77	0.89	0.83	0.82
DNN	0.79	0.85	0.82	0.82
PAC	**0.84**	**0.96**	**0.90**	**0.89**

Fake News: (20800 records)

Model	P	R	F1	A
PAC	0.66	0.88	0.76	0.72
Log reg	0.69	0.84	0.76	0.73
DNN	**0.75**	**0.74**	**0.75**	**0.75**

PAC - Passive Aggressive Classifier
P - Precision
R - Recall
F1- F1 score
A - Accuracy

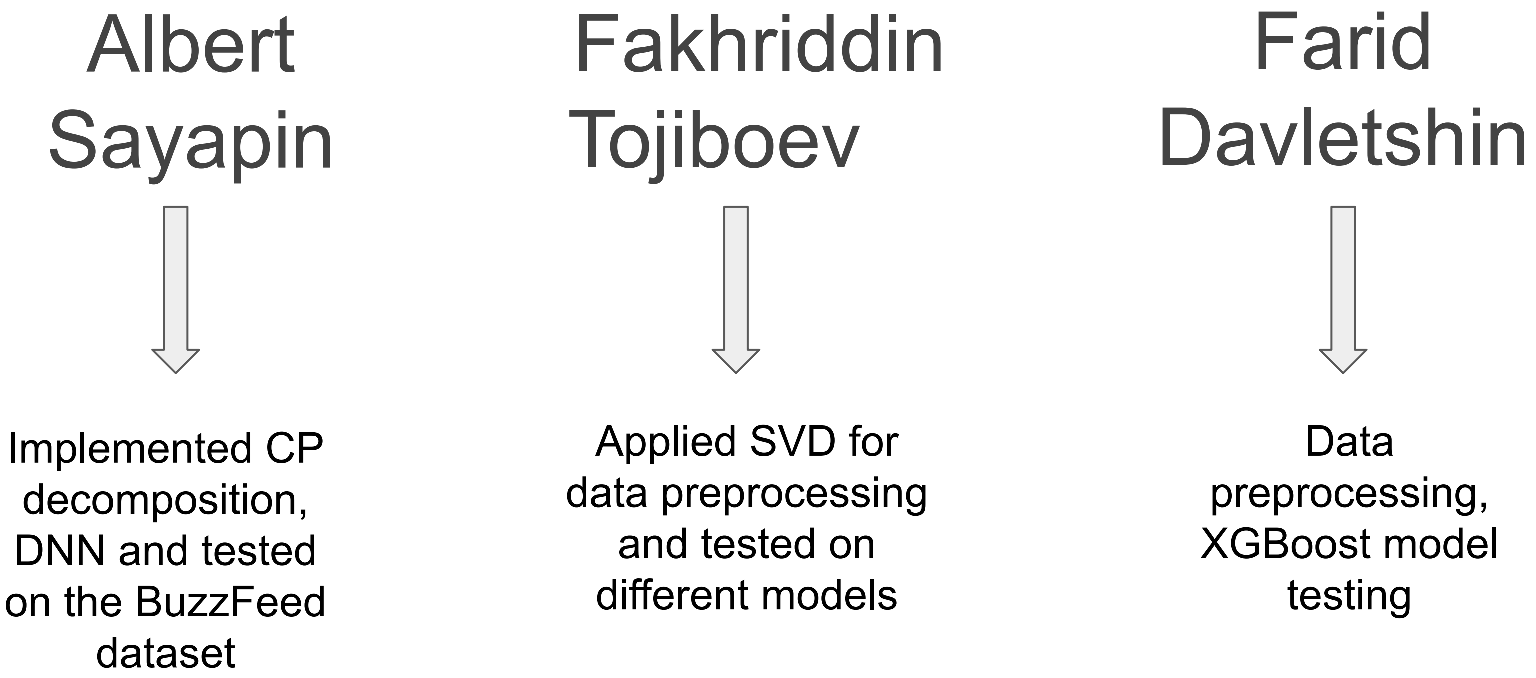
Summary:

- **Main results:** With Tensor/Matrix decomposition we *reduced working on feature engineering* and achieved *better results*
- **What targets were planned to achieve?** To *apply Tensor* and *SVD* decompositions, identify if particular news is fake or real using *features* given *by tensor/matrix decompositions*
- **What targets were not achieved and why?** *All* targets were *achieved*
- **Potential improvement of your project:**

  - Our realization of CP decomposition works well but can be *parallelized* to get results quicker or implemented using cupy to work with *GPU*

  - Check the tensor/svd based methods on the *other datasets*, real-world ones

Summary:



Albert Sayapin

Implemented CP decomposition, DNN and tested on the BuzzFeed dataset

Fakhriddin Tojiboev

Applied SVD for data preprocessing and tested on different models

Farid Davletshin

Data preprocessing, XGBoost model testing

Thanks!

Skoltech

Skolkovo Institute of Science and Technology