Department of Computer Science at the University of Siegen

Institute for Knowledge Based Systems & Knowledge Management

**Exposé for Master thesis**

Web-based application for the detection of fake news using Random Forest and Naïve Bayes

[1 Introduction 1](#_Toc65181442)

[1.1 Problem Statement 1](#_Toc65181443)

[1.2 Course of research and objectives 1](#_Toc65181444)

[1.3 Expected result 3](#_Toc65181445)

[2 Table of Contents (preliminary) for the Master Thesis 4](#_Toc65181446)

[3 So far found and viewed Literature 5](#_Toc65181447)

[4 Work plan with Milestones and Apps Architecture 6](#_Toc65181448)

[5 Publication bibliography 7](#_Toc65181449)

# Introduction

## Problem Statement

The advent of the World Wide Web and the rapid adoption of social media platforms like Facebook, Twitter Instagram other Flickr[[1]](#footnote-1) , paved the way for information dissemination that has not been so often witnessed in the human history before. These social media platforms are extremely powerful and useful for their ability to allow users to discuss and share ideas and debate over issues such as democracy, education, and health. However, this usefulness comes with a backside, because such platforms are also used with negative perspective by certain entities commonly for their own interests. A reader cannot be certain that any newly posted item or information has been thoroughly researched or is intentionally erroneous or misleading. This is a big problem because, we as human beings, our ability to take decision relies mostly on the type of information we consume; our world view is shaped on the basis of information we digest.

## Course of research and objectives

The main purpose of this Master thesis is to develop a web-based application using machine learning (ML) algorithms capable of classifying piece of news into false or true. The classification of these information into “true” and “false” will be supported by an evolved version of two ML models built upon the following algorithms: “Random Forest” (RF) and “Naïve Bayes” (NB). The development of the first version of the ML models has been subject of my previous work (Master Project). The ML models (version 1) have been trained using prelabelled data collected from the ISOT dataset[[2]](#footnote-2) which achieved good classification results both for RF and NB. The evolved ML models version (version 2) which will be used by the web-based application to be developed, extends the version 1 by feeding it with new prelabelled training data gather from website such as PolitiFact[[3]](#footnote-3) using web-scraping. Meanwhile, for the purpose of trying to solve the problem[[4]](#footnote-4) from the root, the author will dive into the motivations for fake news and the psychological reasons for why people believe them (describing why fake news spread so much and how they impact the whole society).

To realize this aim, the steps below are necessary:

* List the characteristics and the motivations behind fake news.
* Give the best and most updated information about what is currently known as fake news. A broad overview of fake news psychology will be provided to answer questions such as: why do human beings fall for fake news? how do fake news spread and why people believe them?
* Provide detailed information on the functioning of “Random Forest” and “Naïve Bayes” and deliver ways how they are used for fake news detection (specially for information coming from social media platforms and fact-checking websites).
* Develop a website based upon machine learning algorithms that will:
* allow users to enter information their want to check. Because the collected data used to train the models are all related to the US politics, between end of 2016 and beginning of 2021, users could only fact-check information in that time frame.
* display the result of the checking,
* allow users to login,
* automatic search for news data through web-scraping to continually train and improve the models.
* Write a documentation that describes all the functionalities of the Web-based application and which cover the tests and test results.

## Expected result

What is expected as final result for this work, is first of all a broad overview of the fake news psychology. Second and finally deliver a website that includes the features described above. Among which automatically train the ML models through data collected from PolitiFact (using web-scraping) and the ISOT dataset. Users should be able to login into the website, to enter information[[5]](#footnote-5) they want to check (whether the latter is true or false) and get the result of classification displayed on the screen.

The technologies that will be used for the development on the website are the following:

* React.js is suggested for the implementation of the front-end features.
* Python, Flask, and a bunch of other libraries will be considered for the implementation of the back-end features, such as “BeautifulSoup” for web scraping.
* A relational database (MySQL or Postgres).
* Docker and nginx for the deployment and load balancing of the different services.

# Table of Contents (preliminary) for the Master Thesis

**I** Table of contents

**II** Table of figures

**III** Table index

**1** Introduction (limit application prediction capacity because of the data used for trainings [date, and type and information])

**1.1** Motivation

**1.2** Problem definition and objectives

**1.3** Methodology

**2** Fake news

**2.1** Definition and Characteristics

**2.2** Why fake news spread, and people believe them?

**2.3** How human can detect fake news

**3** Detection of fake news using machine learning

**3.1** Résumé from previous work

**3.2** Data processing and NLP

**3.3** Random Forest algorithms

**3.4** Naïve Bayes algorithms

**4** Practical part: Implementation of the web-based application

**4.1** Result from previous work

**4.2** Application design, use cases and technical specifications

**4.3** Coding

**4.4** Testing and Documentation

**5** Summary and perspective

**5.1** Summary

**5.2** Perspective

# So far found and viewed Literature

(Ahmad et al. 2020, p. 1)

(Aswini et al. 2018)

(Islam et al. 2020)

(Nicole 2018)

(Amer and Tamanna Siddiqui 2020)

(Barclay 2018)

(Radaideh et al. 2020, p. 2)

(Gerard 2012)

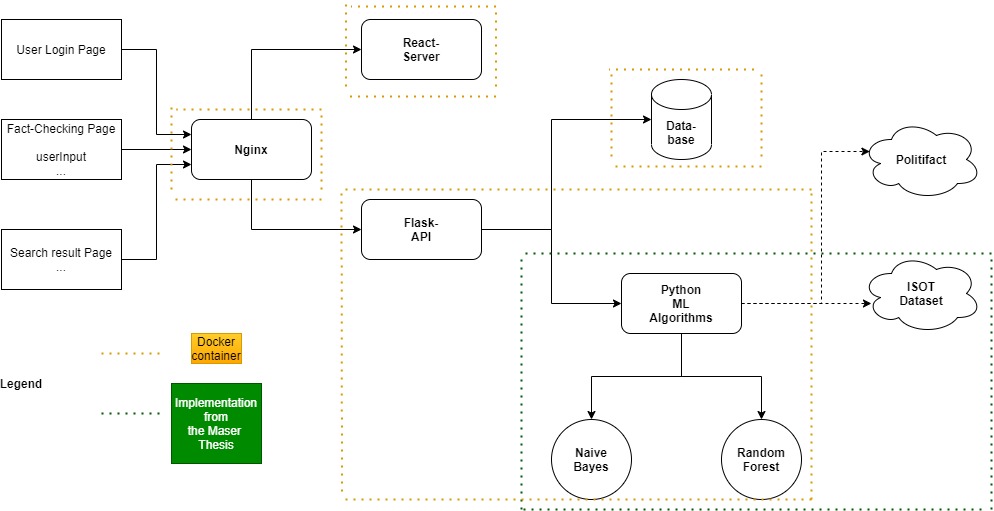
(Vidyarthi et al. 2020, p. 5)

(Yanli et al.)

(Rahat et al. 2019, p. 268)

# Work plan with Milestones and Apps Architecture

A copy of the **Gant-Diagram** will be also attached in the Email.



**Application Design**

Publication bibliography

Ahmad, Iftikhar; Yousaf, Muhammad; Yousaf, Suhail; Ahmad, Muhammad Ovais (2020): Fake News Detection Using Machine Learning Ensemble Methods. In *Complexity* 2020, pp. 1–11. DOI: 10.1155/2020/8885861.

Amer, Abdullah Yahya Abdullah; Tamanna Siddiqui (2020): Detection of Covid-19 Fake News text data using Random Forest and Decision tree Classifiers. DOI: 10.5281/ZENODO.4427205.

Aswini, Thota; Priyanka, Tilak; Simrat, Ahluwalia; Nibrat, Lohia (2018): Fake News Detection: A Deep Learning Approach, checked on 1/31/2021.

Barclay, Donald A. (2018): Fake news, propaganda, and plain old lies. How to find trustworthy information in the digital age. Lanham: Rowman & Littlefield, checked on 1/31/2021.

Gerard, Biau (2012): Analysis of a Random Forests Model, checked on 1/31/2021.

Islam, Md Rafiqul; Liu, Shaowu; Wang, Xianzhi; Xu, Guandong (2020): Deep learning for misinformation detection on online social networks: a survey and new perspectives. In *Social network analysis and mining* 10 (1), p. 82. DOI: 10.1007/s13278-020-00696-x.

Nicole, O'Brien (2018): Machine Learning for Detection of Fake News, checked on 1/31/2021.

Radaideh, Abdulrahman; Dweiri, Fikri; Obaidat, Mohammad (2020): A Novel Approach to Predict the Real Time Sentimental Analysis by Naive Bayes & RNN Algorithm during the COVID Pandemic in UAE. In : 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). Sharjah, United Arab Emirates, 03.11.2020 - 05.11.2020: IEEE, pp. 1–5, checked on 1/31/2021.

Rahat, Abdul Mohaimin; Kahir, Abdul; Masum, Abu Kaisar Mohammad (2019): Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. In Rahat1 Abdul Mohaimin, Abdul Kahir, Mohammad Masum Abu Kaisar (Eds.): 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART). Moradabad, India, 22.11.2019 - 23.11.2019: IEEE, pp. 266–270, checked on 1/31/2021.

Vidyarthi, Sriram K.; Tiwari, Rakhee; Singh, Samrendra K.; Xiao, Hong‐Wei (2020): Prediction of size and mass of pistachio kernels using random Forest machine learning. In *J Food Process Eng* 43 (9). DOI: 10.1111/jfpe.13473.

Yanli, Liu; Yourong, Wang; Jian, Zhang: New Machine Learning Algorithm\_ Random Forest, checked on 1/31/2021.

1. https://www.flickr.com/ [↑](#footnote-ref-1)
2. https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php [↑](#footnote-ref-2)
3. https://www.politifact.com/ [↑](#footnote-ref-3)
4. The fake news problem (creation, spread and people believing them) [↑](#footnote-ref-4)
5. The to check information will be restricted on US politics between end 2016 and begining 2021. [↑](#footnote-ref-5)