MULTIMEDIA UNIVERSITY OF KENYA

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

MULTILINGUAL FAKE NEWS DETECTION SYSTEM

BY

ODHIAMBO PATRICK OLUOCH

CIT-223-040/2017

Submitted in partial fulfillment of the requirements of Bachelor of Science in Computer Science.

DECLARATION

I hereby declare that this project proposal is my own work and has, to the best of my knowledge, not been submitted to any other institution of higher.

Student: _____Registration Number:_____

Signature:_____ Date:_____

This project proposal has been submitted as a partial fulfillment of requirements for the Bachelor of Computer Science of Multimedia University of Kenya with my approval as the University Supervisor.

Supervisor: _____

Signature:_____ Date: _____

ACKNOWLEDGEMENT

I sincerely appreciate my Lecturer MR. PETER MUTURI who sacrificed his time to guide and mentor me to be a critical thinker and solve concrete problems in the society, as a computer science student. He provided a conducive environment for open discussions and this not only improved my communication skills but also made me view problems from various perspectives.

This proposal would not have been successful without the cooperation and support of my aunt Gaudencia Okeyo, friends and other family members who encouraged me never to give up, who funded me in performing my research, and who have promised to always offer support until the end of project execution.

ABSTRACT

There has been a tremendous rise in the spread of fake news, i.e., false information created with the intention of deception. This poses a serious threat to both political, economic and social life, since it fosters political polarization and the distrust of people with respect to their leaders. The overload amount of news that is disseminated through social media makes manual verification tiresome and less accurate since humans are subject to bias, which has promoted the design and implementation of automatic systems for fake news detection. Fake news disseminators use myriads of approaches to promote the success of their creations, with one of them being to excite the stands of the recipients and cause public harm. This has led to sentiment analysis, the part of text analytics concerned with determining the polarity and strength of sentiments expressed in a text, to be used in fake news detection approaches. The previous studies have explained the different uses of sentiment analysis in the detection of fake news. There is need to consider other multimedia elements like images, and different natural languages since multilingualism has not been properly met.

LIST OF ABBREVIATIONS

CPU – Central Processing Unit

GPU – Tensor Processing Unit

SSD – Solid State Drive

HDD –Hard Disk Drive

CI/CD – Continuous Integration/ Continuous Deployment

Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1 INTRODUCTION

1.1 Background study

Fake news existed years before the advent of the internet and the dissemination occurred through the print media such as the magazines, journal and other hard copy productions. The advent of the internet attracted the masses as the universe embraces the concept of digital economy, speed and costs being a consideration.

These Online platforms have always been a cutting-edge sword for news updates in the ever-evolving internet society. On the other hand, online media provides for easy access, little to no cost, and the spread of information at an impressive rate (Shu, Sliva, Wang, Tang, & Liu, 2017). The fake news creators to spread false information can take the superb advantage of faster dissemination. False news has the tendency of spreading faster as compared to genuine information because people tend to base their logic on what they are exposed to but not necessarily, what is true.

Studies have been conducted by scholars with an aim of mitigating the spread of fake news on social media platforms taking into concern the impacts fake information can pose to the society. Fake news can cause harm to individuals and to large cooperation including the government. An example is the impact of fake news on the United States 2016 presidential election on Twitter. The misinformation campaigns altered public opinion and endangered the integrity of the presidential election (Alexandre Bovet & Hernán A. Makse 2019). An example of effects on corporations is when a journalist posted an article on CNN's iReport.com in 2008 that Steve Jobs, CEO of Apple Inc., had a heart attack. Everyone responded to this by sharing the article widely. The caused a fluctuation to the stock of Job's company, Apple Inc. due to a single false news report that had been mistaken for authentic news reporting (Rubin, 2017).

The creation and spread of fake news on online platforms is not only limited to humans but also to programmed social bots, trolls and cyborg users. This has been boosted by the fact that there are less strict solid laws and restrictions against information sharing over the internet. Trolls are real humans who "aim to disrupt online communities" in hopes of provoking social media users into an emotional response (Shu et al., 2017). They do this to confuse the public on the polarity of truth in news items while Cyborg users are a combination of "automated activities with human input" (Shu et al., 2017) owning pseudo accounts and having false identities.

1.2 Problem statement

The spread of fake news has been rampant as compared to the spread of true news particularly political news (Vosoughi, S. Roy, D. Aral, S., 2018). Despite the previous studies conducted by the social media giants and researchers to counter the spread, the issue of multilingualism and the analysis of other multimedia elements other than textual content has not been properly handled. Apart from the use of digital techniques, organizations have partnered with independent fact checkers that greatly rely in human research, which may be subject to bias and inability to handle large amount of work loads, this include Snopes and PolitiFact.

1.3 Proposed solution

My proposal for the solution to the problem under study is a system that intelligently detect the truthfulness in news items in different natural languages and that is less reliant on human intervention. This is based on Natural Language Processing and data mining techniques based on metrics such as author credibility, author-article publication history, information from independent fact checkers like Snopes, Textual content analysis.

1.4 Aim of the study

Develop a multilingual system reliant on optimal sentiment analysis techniques to detect untruthfulness in news items.

1.5 Research objectives

i. Extract data sets on Kaggle about the historical news items and their truth indices.
ii. Detect statistical features and relationships in the dataset.
iii. Develop thoroughly trained and tested model using the cleaned dataset.
iv. Assign truth index to news items as per the level of truth contained herein.

1.6 Significance of the study

The study will aid in the reduction of the rate of spread of false news on social media networks and news outlets hence reduce the negative impacts caused by propagation of faulty and unjustified information by programmed bots and human beings.

1.7 Scope

The study will be constrained to the use of free datasets available on online via Kaggle and through web crawling.

The study will be constrained to the textual data and not images and other graphical formats.

The study is constrained to the use of open source libraries and technologies.

1.8 Assumptions and limitations

The project will be subject to an assumption that news items will be presented as textual content. Therefore, textual data will be used throughout the project from training to testing due to the.

The assumptions poses a limitation to the system in that news is always presented as a composition of multimedia like images, videos and audio.

CHAPTER TWO

2 LITERATURE REVIEW

A look at scholarly articles published indicates clearly that the issue of fake news over the internet is a great topic of concern amongst researchers, and recently has been a great concern in the spread of false news on COVID-19 vaccines. The concern should not only be relegated to the IT department or public relations only, but should be a concern to everyone. Despite the issue of fake news on social platforms gaining more attention in the recent past, still there is no adequate publications to address such. Researchers have proposed various machine learning techniques to identify truthfulness in news items through Natural Language Processing Techniques. Kai Shu et al proposes the use of machine learning ensemble methods to detect faultiness in news items. In a work published in 2019, Kai Shu et al argues that fake news is a complex topic that does not only require a single technique but an optimization of several techniques presented in machine learning as ensemble methods. Xinyi Zhou and Reza Zafarani in an article published in July 2020 proposes the use of knowledge-based, propagation-based, style-based and source-based techniques to detect fake news. The paper classifies knowledge-based as manual fact checking done by human experts and automatic fact checking done through machine learning approaches. Style- based as the analysis being subject to the textual structure and source-based is the deep analysis of the news source credibility by identifying the authors and publication history.

The previous works indicate that fake news imposes integrity as a security implication to data and this creates a challenge in the current business world where data is an asset. Wrong data implies wrong forecasting plus poor results.

2.1 Related systems

Google fact check tool

This is a fact checker by google that retrieves information about a particular statement, from sources on the web and displays to the user depending on the query keyed in by the user. The site is steered by the keyword search and does not give the level of truthfulness on information feed in but gives the various instances where similar statements appear on the web.

PolitiFact Website

This is a website platform run by editors and reporters from Tampa Bay Newspaper. It is majorly detect veracity on political news, with geographical span being America and it is environs. The researchers and reporters perform intensive research on news and assign indices based on the level of truth on the information.

The indices (PolitiFact O meter) include True, mostly true, half true, mostly false and pants on fire.

PolitiFact is more accurate, however, the fact that it is much reliant on human researchers makes it least considered in the analysis of bulk news posts from social platforms and news outlets before being posted to the public.

Snopes Website

Similar to PolitiFact, the fact checking process is much dependent on human researchers, there is less automation, and the geographical constrain is America and its environments.

2.2 Limitations of the existing systems

I. Google fact check tool though is automated and is less dependent on human, is based on checking for facts on the metrics of how frequent is it on other sites across the web.
II. Snopes, PolitiFact, and others like factcheck.org systems implement the idea of fact-checking journalism to identify facts on news.
III. Both Snopes and PolitiFact are geographically limited to American region hence does not fully accommodate other regions and languages.
IV. Human beings are subject to bias and gets overwhelmed with much loads of data.

2.3 How the proposed system solves the challenges

I. The proposed system will support different languages other than English language.
II. The proposed system will rely much on optimal machine learning algorithms as opposed to human researchers.

3 METHODOLOGY

3.1 Introduction

The system will make use of the different Natural Language Processing Techniques to come up with truth-values to news items and reduce the level of bias posed by human fact checkers by media institutions.

The solution to the problem will also rely on the use of web crawlers in data mining across web pages, language translators and News Outlets Application Programming Interfaces to extract data items for analysis.

The detection of truth in news items will involve various metrics which are not limited to; subject credibility analysis which trains a model on truth values of different news subjects, creator-article publication history(Number of articles an author has in history), articles credibility with textual content analysis, creator credibility analysis.

The implementation will be based on but not limited to data mining and machine learning classification techniques.

The proposed techniques are:

Classification techniques

Classification technique will be used to assign a truth-value to news items after performing various sentiment analysis.

Classification algorithms applied on training data will be used to detect patterns in data and assign a label based on whether an item is true or false as per the defined truth meter.

i. **Naïve Bayes**

   This technique calculates the possibility of whether a data point belongs within a particular category or does not. We will be using these techniques to categorize words and phrases as belonging to a preset tag or not.

   For example with naïve Bayes, we can check whether a news item is false, partially false, true or partially true.

This employs the concept of probability in its implementation for instance; we may test for the probability of a news being true when the author's credibility is guaranteed. Bayes has a general equation of:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

This implies the probability of A, if B is true, is equal to the probability of B , if A is true, times the probability of A being true, divided by the probability of B being true.

ii. Decision Tree

We will use decision trees to categories sentences into phrases phrases into words and make intelligent decisions on the tree. This will help us to create categories within categories, allowing for organic classification under limited human supervision.

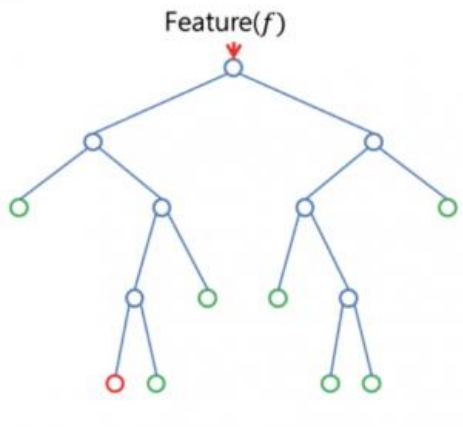

Figure 1Decision Tree

iii. Random Forest

The constructed decision trees will be ensemble to obtain a final tree to obtain a more accurate and stable prediction because additional randomness is achieved while growing the tree.

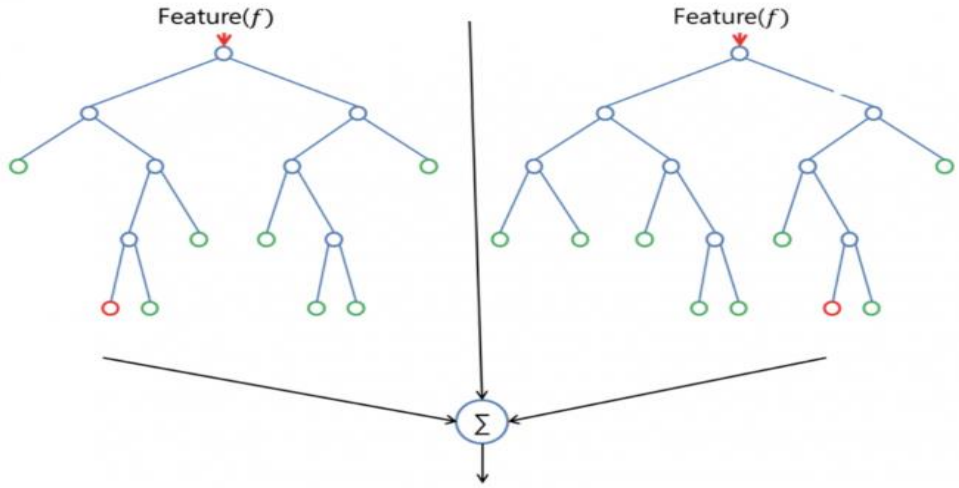

Figure 2Random Forest

iv. Support Vector Machines

This technique will be used in training and classifying news items within degree of polarity. This helps in identifying and categorizing various news items as true or false.

Through this we will be able to analyses various sentiment lexicons and assign appropriate labels to them.

Justification

It is justifiable to choose the above techniques since they offer the best sentiment analysis machine learning approaches, in terms of accuracy and performance. Support Vector machines can both be used in supervised learning to classify data items into categories and unsupervised learning too.

3.2 Agile Development Methodology

These methodologies are rooted in adaptive planning, early delivery and continuous improvement, all with an eye toward being able to respond to change quickly and easily.

However, as more and more development teams adopt an agile philosophy, testers have struggle to keep pace. That is because the widespread adoption of Agile has led teams to issue releases and much undocumented software on a more frequent basis. This frequency forces testers to shift when they conduct testing, how they work with developers and even what tests they conduct, all while maintaining quality standards.

In order to achieve better of an agile development methodology, the methodology will be incorporated in a CI/CD pipeline to ensure that testing and development do not overlap.

3.2.1 Reasons for Choosing agile development methodology

I. Improved quality: by adapting this methodology, organizations can deliver; organization can deliver solutions in time and with the higher degree of client and customer satisfaction.
II. Focus on business value through increased focus on delivering strategic value by involving business stakeholders in the development process.
III. Focus on users: agile development methodology uses user's stories with business focus acceptance criteria to define product features.

IV. Stakeholder engagement: this provide more opportunity for the team to truly understand the business vision, deliver working software early and frequently increase stakeholders trust.

V. Transparency: this can include prioritizing features, iteration planning and review section or frequent software builds containing new features

3.2.2 Challenges of agile development methodology

I. People's behavioral change: changing the way people work is difficult- the habit and culture of large development organization are typically in grain. People naturally restrict change and therefor when confronted with an agile transformation.

II. Lack of skilled product owners from the business side: most product owners do not understand user stories and hesitate to give up the BRD for something different because they view it as a contract between them and IT.

III. Lack of dedicated cross-functional team: In most cases, there has always been inefficient cross-functional team.

3.3 DATA COLLECTION METHODS

3.3.1 Questionnaires

This was done through eliciting the feelings, beliefs, experiences, perceptions or attitudes of some members of the public. This was conducted through online forms via Google forms and Microsoft forms. Direct issues to questionnaires to individuals supplemented these online forms. As a data collection instrument a questionnaire can be *structured, unstructured* or *semi structured.*

A structured Questionnaire is one that has closed ended questions. It is restricted and calls for a "yes" or "no" answer.

Unstructured questionnaire is one that has open-ended questions. It is unrestricted and calls free response from the respondent. Semi structured questionnaire has both open and closed ended questions.

Advantages

I. It has a low costs-even when the universe is large and widely spread geographically, google forms are free of change easy to design and distribute via social media.
II. Online questionnaires are relatively free from bias of the interviewer.
III. Uniformity of the questions i.e. standardized questions.
IV. Respondents have adequate time to give all the answers in the convenience of their time.

Disadvantages

I. It has a low rate of return.
II. Respondent's motivation is difficult to assess since there is less physical interaction with individuals.
III. May present biased samples
IV. It can only be used when respondents are educated and cooperative. The control of the questionnaire may be lost once it is sent.

3.3.2 Interviews

It involves presentation of oral-verbal stimuli in terms of oral responses. This method can be used through personal/ telephone interviews. Personal interviews involve an interviewer asking the respondent questions in a face-to-face contact. It is a conversation in which the roles of the interviewer and the respondent change continually. They may be structured interviews where a guiding questionnaire (interview schedule) is used or unstructured interview where there is no questionnaire to be followed. Structured interviews are rigidly standardized and formal while unstructured interviews are flexible and informal

Advantages

i. Helps a researcher to get more information and in greater depth.

ii. It can also be applied to record verbal answers to various questions.
iii. Sample can be controlled.
iv. Can be used with young children and illiterates
v. Allows the interviewer to clarify questions
vi. The language of the interviewer can be adapted to the nature of the respondent
vii. The interviewer can collect supplementary information.

Disadvantages

i. It is expensive to achieve.

ii. Gaining access to interviewers may be very difficult especially if they are high profile people 3. It is time consuming.

3.3.3 Observation

It is commonly used in studies related to behavioral science. It has to be systematically planned and controlled and subjected to checks and controls on validity and reliability and constructed to serve a formulated research purpose for it to serve as a scientific tool for data collection. Direct observation is a measuring instrument to measure such traits as self-control, cooperativeness, truthfulness and honesty. One observes without asking questions to correspondence.

Advantages

1. The researcher is enabled to record the natural behavior.

2. It is done in a natural behavior thus, much bias is reduced.

3. Observation is relatively cheap.

4. It allows collection of a wide range of information

5. It is ideal in studying non-verbal communication.

Disadvantages

1. Observation lacks control of variables in its natural set up.

2. There is difficulty in quantification because it is mostly descriptive.

3. It lacks privacy and has limited study.

Observation studies use a smaller sample than survey studies.

3.3.4 Secondary sources

Data obtained via secondary sources include web sources, journals, textbooks, and eBooks and research papers.

Advantages

1. It is economical. It saves efforts and expenses.
2. It is time saving.
3. It helps to make primary data collection more specific since with the help of secondary data, we are able to make out what are the gaps and deficiencies and what additional information needs to be collected.
4. It helps to improve the understanding of the problem.
5. It provides a basis for comparison for the data that is collected by the researcher.

Disadvantages

1. Secondary data is something that seldom fits in the framework of the marketing research factors.
2. Accuracy of secondary data is not known and can rarely be justified.
3. Data may be outdated.

3.4 Project resources

For the project to achieve its stated objectives from the analysis to deployment phase there are a few tools that will be required. These tools range from software to hardware.

3.4.1 Hardware resources.

Personal computer/Laptop

This is the hardware that the source code of the project will run on and which the testing will be performed.

Since the machine learning algorithms may be so much compute intensive, a laptop with the following specifications is preferred:

- At least 4G RAM computer.

- Internal storage of relatively 500GB HDD or SDD to host the other software needed to implement the system.
- Graphics Processing Unit or a Tensor Processing Unit of global memory access of up to 48GB, around 200cycles and shared memory of about 164kb.

High speed and secure server, with higher latency for securely deploying the API.

3.4.2 Software resources

An integrated development environment PyCharm Community Edition that is freely available for developer community.

Google colabs – this is freemium platform by google which has the necessary hardware to accommodate highly compute intensive deep learning algorithm.

Scikit-learn – this is an open source Python machine-learning framework.

Scrapy – this is a python web crawler. We will be using this in implementing our web crawlers to perform data mining on web pages. It is an open source library.

3.5 Project Schedule

Project schedule

TASK DESCRIPTIC	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
CHAPTER 1 AND CHAPTER 2					
Background	1	6	1	6	100%
Problem study	2	4	2	4	100%
Solution draft	4	6	4	6	100%
Justification	6	6	6	6	100%
Related systems	6	7	6	7	100%
CHAPTER THREE					
Methodologies	7	8	7	8	100%
Data collection	8	9	8	9	100%
Project Schedule	8	9	8	9	100%
Project Budget	9	9	9	9	100%
CHAPTER FOUR					
System Analysis	9	10			0%
CHAPTER FIVE					
System Design	10	11			0%
CHAPTER SIX					
system implimentation	11	12			0%
System Testing	11	12			0%

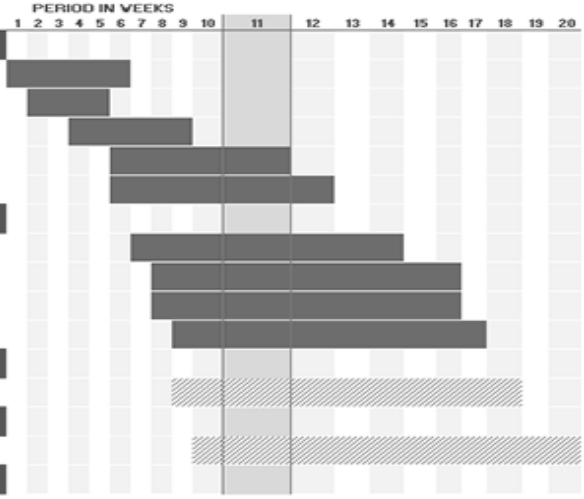

Figure 3 Project schedule

3.6 Project budget

Below is the projected budget of the project.

Name	Description	Amount required
High specifications personal computer.	System development and testing will be done from this computer.	Ksh. 150,000
Research expenses	This include subscriptions to online libraries and learning websites.	Ksh. 100,000
Professional services	This includes the legal pieces of advice from professional and grants to access various news media resources.	Ksh. 150,000
Contingency reserves	This will allow for flexibility and reduce the risk of budget overruns	Ksh. 200,000
Travelling expenses	This is the capital set aside to accommodate for travelling while doing field research	Ksh. 100,000
Hosting services fund allocation	The API will be hosted in a secure and high-speed server that can accommodate higher traffics with higher latency.	Ksh. 200,000 (starting cost)
TOTAL		Ksh. 900,000

Table 1Project budget plan

References

Vosoughi, S.; Roy, D.; Aral, S. (2018) The spread of true and false news online. *Science* , *359*,

1146–1151.

Kesavaraj, G ; Sukumaran, S (2013) A study in classification techniques in data mining. *Data Science*

Soon, W; Ng, Hwee; Lim, D. (2017) A Machine Learning Approach to Coreference Resolution of Noun Phrases.*Computational Linguistics*

Beysolow , T.I. (2018). Applied Natural Language Processing With Python Apress

Secker ,Davies, Freitas, Timmis, Mendao. (2019) An Experimental Comparison of Classification Algorithms for Hierarchical Prediction of Protein Function.
http://www.cs.kent.ac.uk/projects/biasprofs