Bilkent University

Department of Computer Engineering

Senior Design Project

Newspector: Unreliable News Detection

Project Specifications Report

Ahmet Ayrancıoğlu   21601206
Deniz Dalkılıç   21601896
Kaan Gönç   21602670

Supervisor: Varol Akman

Jury Members: Halil Altay Güvenir and Özcan Öztürk

1.   Introduction

People are concerned about the presence of fake news [1] all around the world for several reasons. The term "Fake News" is interpreted as "false news stories, often of a sensational nature, created to be widely shared or distributed for the purpose of generating revenue, or promoting or discrediting a public figure, political movement, company" , according to the Oxford Dictionary. Spreading of fake news is a real and extremely serious threat in today's society as it can lead to violations of rights to information and expression. This in return can be used as a tool of manipulation in terms of prejudice, marginalization, excessive assimilation of ideas, acts, people, products or organizations. Furthermore, there are excessive amounts of fake news present on the internet. The main reasons for this are the ease of writing/authoring fake news; newspapers, magazines, and news agencies having their online channels, and social media being an effective medium for transmission. The internet and social media have made it very easy for anyone to publish content on a website, blog or social media profile and potentially reach large audiences. As most of the people who are reading news is getting the material from social media, many content creators use this as an advantage. These kinds of news can be encountered daily in various fields such as politics, economics, social life, security, humanity. Most importantly, it is extremely hard to prevent them for 3 main reasons. Firstly, it is difficult to identify the source of the news. Secondly, it takes too much human effort and time to crosscheck and confirm their reliability. Finally, they instantly start to spread around the internet when they are published.

Even though there is a colossal amount of fake news on the internet and they may seem random and unrelated in their nature of origin, after analyzing fake news, we start to see a pattern emerge and we can roughly categorize fake news into six. These categories are clickbait, propaganda, satire, sloppy journalism, misleading headlines, and biased news.

Clickbait

Clickbaits are "material put on the Internet in order to attract attention and encourage visitors to click on a link to a particular web page". Clickbait stories use sensationalist headlines to grab attention and drive click-throughs to the publisher's website, normally at the expense of truth or accuracy [2].

Propaganda

Propagandas are ideas or statements that may be false or exaggerated and that are used in order to gain support for a political leader, party, etc. [2].

Satire

Satire is a way of criticizing a person, an idea or an institution in which you use humor to show their faults or weaknesses; a piece of writing that uses >this type of criticism [2].

Sloppy Journalism

Sometimes reporters or journalists may publish a news story with unreliable information or without checking all of the facts which can in return mislead audiences [2].

Misleading Headlines

Stories that are not completely false can be distorted using misleading or sensationalist headlines. These types of news can spread quickly on social media sites where only headlines and small snippets of the full article are displayed on audience newsfeeds [2].

Biased News

Many people are drawn to news or stories that confirm their own beliefs or biases and fake news can prey on these biases. Social media news feeds tend to display news stories and articles that they think the reader will like based on gathered personal data about the reader's daily activities and searches [2].

For the final thing, additionally to the topic of fake news, when we surf on the internet we are exposed to news, articles and other content based on our own routines on the web. Such content tends to reflect our own likes, views, and beliefs and therefore isolating us from differing views and opinions present on the web. Therefore, people do not realize how beneficial it would be to access contents from different perspectives which would highly increase the ability of critical thinking. In addition to identifying the presence of fake news, offering clusters of news, which approach differently to a common topic, to the users is another important issue in this concept. Such a solution that defines this approach as its main methodology would provide many sources for the user to interpret and decide which source or sources are more reliable than others.[3]

Our Mission

According to media personality Hugh Linehan, "Media is no longer passively consumed – it's created, shared, liked, commented on, attacked and defended in all sorts of different ways by hundreds of millions of people. And the algorithms used by the most powerful tech companies – Google and Facebook in particular – are brilliantly designed to personalize and tailor these services to each user's profile." As we particularly focus on the topic of fake news and how to increase the awareness of people against such content, we should first look at how global companies approach this problem. Leading technology companies like Google and Facebook have announced how to identify fake news with reporting and flagging tools. Furthermore, some media companies like BBC and Channel 4 are known to have their own fact-checking sites. Even though these types of movements exist against fake news, all of these processes mainly proceed manually, and therefore it takes too much time and effort to verify or deny such news. The ways for spotting fake news can be divided into different tasks:

- Checking the source of the news
- Checking your own biases
- Checking other sources for discrepancies
- Checking whether the article is a satire or not
- Reading the whole article as well as the headline
- Verifying the facts inside the article

According to our research, there is not a solid tool which successfully automates these processes of fake news detection on the market by using the linguistics analysis between multiple platforms, and we, as a team, believe that all people who encounter various news on the internet need such a tool which will interactively alert them on the unreliability factors of such news. Therefore, our mission is to design and implement such a tool and observe its impact on people's approach to the presence of unreliable news in daily life, especially on the internet.

1.1. Description

The main goal of the *Newspector* is to caution readers about possible unreliability and deception in the news. In order to achieve this, Newspector examines the news articles that are currently being read by the user as the user reads them and scans the other credible sources, tries to find the same or very similar claims with the original article and compares them. When it encounters any signs of a contradiction, the automated system interacts with the reader by a warning that contains the problematic claim, the reason for suspicion and the sources of the contrary claims.

Newspector will use Natural Language Processing, Text Analysis, Sentimental Analysis algorithms to examine and make comparisons of news articles.

The platform will consist of three layers. The first one will be a web browser plugin that the user installs. The other one will be a web page that the user is directed to. The final one is the main server that executes all the examination algorithms, analysis and comparison processes mentioned before. After the plugin extracts the article and sends it to the server, the server will make its computations and will respond back to the plugin with the possible unreliabilities. Regarding the result coming from the server, the plugin will highlight the claims that may be unreliable in the article and will show a popup that asks the user to follow a path to see the contradicted or supported articles on the Newspector web page where we will also put advertisements to support the development process of the Newspector.

1.2. Constraints

1.2.1. Implementation Constraint

For the development process of the NLP and Machine Learning algorithms of the Newspector, we will use the Python programming language. The frontend application will be implemented using HTML and CSS. For data management, our plan is to use Google's Firebase platform. The development process will be controlled and managed by the git version control system.

1.2.2. Economical Constraint

The Newspector is a free-to-use platform and does not have a sign-up system. So the implementation process should assess the minimum amount of expense for technical tools, libraries, APIs, data samples, etc. On the other side, Newspector should try to maximize the amount of income from advertisements without losing user engagement. Most importantly, as our main goal is to provide a tool to raise awareness on the presence of fake news for each person without regarding their economic status, our plan is not to implement a premium membership system but rather use advertisements to make a profit.

1.2.3. Political Constraint

Since the Newspector will label some news articles as "possibly unreliable" if necessary, the user should not feel that these warnings are biased in a political manner such as supporting a specific political party or view. To achieve this feeling of objectiveness, sources of articles that will be examined to cross-check "possible unreliable" articles should be selected carefully and with no bias. For example, if two sources would be selected from one political view, two more sources should be selected from the opposing view.

1.2.4. Social Constraint

Similarly to Section 1.2.2, The Newspector should not make the user feel like the warnings raised by the platform are biased in some social manner. For instance, the users should not feel like articles marked as "possibly unreliable" are biased towards a specific ethnicity, religion or social class.

1.2.5. Privacy Constraint

Since personal data protection is an important issue and there are regulations about it such as General Data Protection Regulation (GDPR) by European Union (EU), the Newspector won't store or use any data related to a user. Some data can be gathered for the sake of enhancing the current system. For instance, to broaden the datasets of the machine learning algorithms used. However, none of the gathered data should be associated with a user.

1.3. Professional and Ethical Issues

Developing software as a group is a challenging task on its own and continuing this development process without encountering a significant problem in a year-long project makes it even more challenging. To achieve an untroubled development process, communication between group members becomes key. Each member has to act like professionals, treat each other as equal and be aware of the responsibilities that they have to each other. Even though each member has their own schedule and other responsibilities, they have to equally contribute to the project and manage their time to do so.

Newspector will be a browser plugin that will mark the website the user is on. To do so, the plugin has to scrape and read the HTML of the website. Web Scraping can cause legal problems and the website being scraped can sometimes block access if it detects a bot is extracting the data. However, web scraping will only be used for the single page the user is reading and the information gathered from that website will not be used against the interest of the original owner of the data.

Data such as what suggested news articles are clicked or what warnings the user interacts with can be gathered and stored to improve the usability and the effectiveness of the Newspector. However, this information will not be collected without the consent of the users and if the user gives their consent for data gathering, the collected data won't be associated with users directly and will be kept anonymously to increase the privacy of users. Additionally, ant information that will be collected will not be shared with any 3rd parties.

Any software, framework, algorithm, etc. that will be used in the development of the Newspector will be licensed and credited accordingly. Additionally, we will try to use open-source software as much as we can to support and contribute to the open-source community.

Finally, because Newspector aims to increase the awareness of the newsreaders and try to reduce bias and prejudice by suggesting other perspectives and sources for the user to read, there can't be any bias in the program. The developers of the Newspector cannot accept any incentives to favor one newspaper or the other.

2. Requirements
2.1. Functional Requirements

Website
- The Newspector website should categorize news into different custom categories.

- The Newspector website should provide information about the news resources that are being used to verify different claims.

- The users should be able to search for news on desired topics.

- The users should be able to navigate to the source website of the news currently being examined in the Newspector website by clicking a button.

Plug-in
- The examination and analysis process should start working when the plug-in application is added to the web browser without needing any additional setup.

- The plug-in application should provide warnings on the user's current web page, highlighting the "unreliable" claim.

- The users should be able to navigate to The Newspector website through the link provided in the pop-up notification.

Server / Database
- The Newspector should be able to update its news database frequently from various sources.

- The daily activities of users should be used to improve the dataset of the Newspector.

- The Newspector website should support advertisements, as the plug-in will not include any ads for user experience.

- The news database should be backed up regularly.

- The Newspector database should store articles under related categories.

2.2. Non-Functional Requirements

2.2.1. Supportability

- All classes and scripts should have a description explaining the functionality of the class and the reasoning behind its creation. Additionally, all functions should also have brief descriptions that give insight into why the developer chose to write that function.

- The project should be divided into well-defined and independent subprojects and layers in order for multiple people to work on the overall application together efficiently.

- The project should be structured in a way that would support extending the project even though some of the decisions for achieving said extendibility could be considered over-engineering by some for the initial state of the project.

2.2.2. Usability

- Since Newspector will be a browser plugin and will show warnings over the news, the warning should be seen with ease but should not disturb the user from reading the news article.

- The warning popup should be able to clearly show the user where the unreliability resides and should provide a button for the user to go the Newspector website for further analysis of the article.

- The Newspector should be available to download in the Chrome Web Store for ease of access.

- The plugins should have an option to show users a page with instructions and a rough explanation of how the Newspector works behind the scenes.

2.2.3. Reliability

- Gathering the news pieces from selected sources is crucial for the Newspector to function. So, Newspector should be able to gather news from at least 90% of the selected sources at a given time.

- To increase the trust of users towards the Newspector, false reporting of solid well supported information that can be validated by other sources should be kept at a minimum of 10%.

- The sources should be selected carefully by a knowledgeable person in order to reduce bias.

- Newspector should update its sources so that no deadlinks will ever be presented to a user such as removed news articles, etc.

- The data storage solution for the project should be selected in a way that will allow the Newspector to be functional when a crash happens on one of the servers that store the news data.

2.2.4. Performance
- The gathering of the news from selected sources should be fast and effective to include the most up to date news for cross-validation.

- Examination of the news article being read by the user should be fast and be done in real-time as the user reads.

- The Newspector plugin should not cause any slowdown on the website the user is on.

- The servers running the cross-validation algorithms should be fast and can be auto-scaled as the population of the userbase grows to provide the same performance for every user.

2.2.5. Extendibility
- The addition of further news sources should be easy. Other parts of the application should not be dependent on where the news is coming from.

- The addition of further categories should not affect the rest of the cross-validation algorithms.

2.2.6. Portability
- The process of porting the Newspector to mobile platforms should be easy and should be connected to the same backend server.

- The user experience of the Newspector should be the same on various platforms.

2.2.7. Privacy

- The use of cookies and other means of gathering information should be done with the permission of the user.

- Data gathered by users should not be associated with users directly unless they give their consent.

- Data gathered should be securely stored in our databases as it can easily be used to target news or advertisement by other platforms.

3.   References

[1] McCarthy, Niall. "Where Concern Is Highest About Fake News On The Internet [Infographic]." *Forbes*, Forbes Magazine, 12 June 2019, https://www.forbes.com/sites/niallmccarthy/2019/06/12/where-concern-is-highest-about-fake-news-on-the-internet-infographic/amp/.

[2] "Oxford Learner's Dictionaries: Find Definitions, Translations, and Grammar Explanations at Oxford Learner's Dictionaries." *Oxford Learner's Dictionaries | Find Definitions, Translations, and Grammar Explanations at Oxford Learner's Dictionaries*, https://www.oxfordlearnersdictionaries.com/.

[3] Métais Elisabeth, et al. *Natural Language Processing and Information Systems: 24th International Conference on Applications of Natural Language to Information Systems, NLDB 2019, Salford, UK, June 26-28, 2019, Proceedings*. Springer International Publishing, 2019.