An Assessment of ML-based Sentiment Analysis for Intelligent Web Filtering

Nearchos Paspallis
npaspallis@uclan.ac.uk
University of Central Lancashire–UCLan Cyprus
Larnaca, Cyprus

Panayiotis Panayiotou
p.panayiotou@i-spiral.com
Complytek
Larnaca, Cyprus

ABSTRACT

Many people are increasingly using the Web both to accomplish tasks but also for entertainment. At the same time, for many people a significant fraction of their web-browsing time is spent on reading news, both from social media sources and traditional news outlets. However, often this comes with a mental price when facing negative news: research shows that *doomscrolling* can have a negative impact on your well-being. In this paper, we discuss the development and evaluation of an intelligent web filtering mechanism, in the form of a web browser extension (*i.e.*, plugin). This mechanism aims at providing an intelligent assistant that filters out undesired content, aiming at improving the user's well-being. The effectiveness of this approach is assessed with an evaluation involving 50 participants. Our results show that our approach is acceptable by the participants and that there is indeed a need for such tools.

KEYWORDS

intelligent filtering, well-being, sentiment analysis, machine learning

ACM Reference Format:
Nearchos Paspallis and Panayiotis Panayiotou. 2024. An Assessment of ML-based Sentiment Analysis for Intelligent Web Filtering. In *The PErvasive Technologies Related to Assistive Environments (PETRA) conference (PETRA '24), June 26–28, 2024, Crete, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3652037.3652039

1 INTRODUCTION

The Web is praised as one of the most significant technological advancements of the last decades. A majority of the world population uses it daily for performing everyday tasks (such as using email, calendar, etc.) and for entertainment (reading news, watching videos, etc.) News outlets are one of the main forms of informing people of the latest developments locally, regionally, and internationally. In this sense, it is not surprisingly that Web news outlets have gradually surpassed their printed counterparts.

Often, media outlets share negative news. This is due to the fact that they are to inform the public of real, negative events happening on the planet (accidents, wars, etc.) And, to some extent, because negative news can attract the interest of a wider viewership compared to the more neutral or positive news. People are often paying more attention to negative news as they naturally aim to understand whether these pose a threat for their person or for their circle. For example, the Covid-19 pandemic gave rise to *doomscrolling*, defined as "*an excessive amount of screen time devoted to the absorption of dystopian news*" [2].

The impact negative web media content has on its users can not be ignored. The long-term effects of the continuous consumption of irrelevant and unwanted Internet content may not be visible at first but are inevitable. The majority of Internet users spend a substantial fraction of their time browsing news and social media sites, without realizing the physiological and sometimes physical consequences of their simple yet impactful action of endless scrolling [4, 7].

The community has noticed this issue and many have tried to develop tools to limit access to negative web media content consumed by Internet users, but with limited success. Because of their low adoption, such tools have not had a significant impact helped to limiting the negative effects of the unwanted and irrelevant information presented throughout the web. It has been argued that this also contributes to a *digital divide* [22].

Large companies which possess the required resources to develop impactful tools prefer to focus on projects that can generate revenue without high risks. These projects mostly target other companies and their need to generate actionable data from their customers and online products through the use of AI. As a result, such projects often ignore the need of individuals to control and limit the amount of irrelevant information presented to them while Web browsing.

Lastly, while a large number of users prefer to browse the Web via their mobile phones or tablets, there is still a large percentage of users that use a desktop or laptop computer for at least some of their Web browsing. This includes browsing for both work and recreation. With this in mind, we propose a web browser extension which could filter out unwanted and irrelevant items in real-time.

This paper discusses the design and development of a Web browser extension which can identify unwanted or negative content on the Web, and hint to the user to choose to view it or not. To build this extension, we evaluate existing ML (*Machine Learning*) algorithms, and assess their effectiveness in terms of correctly identifying the sentiment of given text. For its evaluation we compare open datasets which are used for training the selected sentiment analysis algorithms, and selecting the most suitable for the purposes of the browser extension. To facilitate its evaluation with real users, we provide a testing framework and mock Web pages to test and verify the functionality of the Web browser extension.

The main achievements of this paper are:

PETRA '24, June 26–28, 2024, Crete, Greece

ACM ISBN 979-8-4007-1760-4/24/06
https://doi.org/10.1145/3652037.3652039

- We develop and present a prototype web-browser extension which uses tested and proven algorithms to detect the *sentiment* of the analysed articles.
- We assess the need for such an extension, as well as its usefulness and performance, via an evaluation which involves real users testing the prototype.

The rest of this paper is organised as follows: Section 2 reviews the related work and provides the context for this paper. Then Section 3 discusses the design and implementation of the browser extension (*i.e.*, plugin), as well as the evaluation and selection of the underlying ML datasets and algorithms used for realising the filtering. Our approach is assessed by means of a user evaluation, as described in Section 4, and the results are further analysed in a discussion presented in Section 5. We close the paper with the conclusions and directions for future work in Section 6.

2 RELATED WORK

Artificial Intelligence (AI) and Machine Learning-based approaches have had a great impact on the world, especially over the last two decades. From autonomous driving [37] to assisting programmers with automatic code generation [8, 15, 23, 36], AI is rapidly transforming the fabric of our modern world. The need for, and the wide applicability of ML algorithms have even prompted the development of specialised processors, aiming to support ML algorithms efficiently and effectively [28, 29]. More recently, the application of LLMs (*Large Language Models*) especially in Chatbots such as ChatGPT, has gained significant publicity for their capabilities [14, 17] and also triggered waves of reflection on how this could impact the world from an ethical perspective [33].

At the same time, the unprecedented advent of information technology, has resulted to most humans spending a significant fraction of their time browsing the Web. This has raised some concerns, *e.g.* with risks related to the continuous consumption of Web media content. Some risks are well known and have triggered the introduction of regulation, with some success [9]. However, some of these risks are only now identified as significant. One such risk is the unconscious consumption of *negative* web media content and its negative impact on mental health [12]. The unwanted or irrelevant information presented to users on the Web can have a major impact on their mental and physical state, and can degrade their general well-being.

In the past, research from Ahern *et al.* [1] which focused on the impact of graphic television images found that these "*may exacerbate psychological symptoms in disaster situations.*" Similarly, research from Boukes and Vliegenthart [3] showed that "*the consumption of hard news television programs has a negative effect on the development of mental well-being over time.*"

In recent years, the COVID-19 pandemic has had a profound impact on people all over the world, even causing changes in behaviour, including to the rate with which we consume media. This has prompted many researchers to assess the impact of extended exposure to news on the viewers' well-being. For example, Zahava *et al.* [31] have identified "*the negative implications of TV news watching during a mass trauma for traumatized individuals*". Similarly, the authors of [24] have further showcased the impact of news on well-being by showing that "*Perceived vulnerability to COVID-19 can serve as a pathway through which exposure to COVID-19 news on mainstream media may be associated with depressive symptoms.*"

Web filtering is not new. Often, web filtering is applied at the network level, where certain domains may be blocked if deemed inappropriate or unproductive for a particular cause. For example adult websites may be blocked from kids' computers, and social media sites may be blocked in work environments. On the other hand, developments in web browser technology, and especially the adoption of a plugin architecture in most modern web browsers, allows for approaches that apply the filtering in the web browser itself. For example, the authors of [34] have proposed "*a browser extension that can effectively and accurately filter profanity, bad words and words with double meaning.*".

In the specific domain of mitigating the impact of exposure to negative news, there are very few tools available. Such tools aim to minimise the impact of negative news by monitoring and hiding such information before it is presented to the user. For example, the *Detox Browser* [21] provides an automated web filtering tool which aims to improve the users' well-being, similar to our proposal. This approach performs a sentiment analysis of the title and first lines of text of the article which minimises resource usage but, on the other hand, can lead to less accurate predictions.

3 DESIGN AND IMPLEMENTATION

The implementation of the proposed Web browser plug-in consists of two main components: First the *front-end*, developed in *jQuery* which realises a very basic control panel allowing the user to monitor and configure the plug-in. For instance, the user can opt-out from filtering in selected websites, or can turn off the extension altogether. Additionally, the user can monitor the total number of items processed, and the analysis result (*i.e.*, how many were classified as *positive*, as *neutral*, and as *negative*).

The *front-end* also provides functionality for applying a blur mask on the filtered items, as shown in Figure 1. While the extension will explicitly blur all items that it assesses as *negative*, the users remain in the loop as they can easily use a button to reveal the hidden text and images.

To facilitate the development of the extension, the actual ML-based filtering is placed in a *back-end*, realised with Python and Flask. The *back-end* provides an interface where individual text-based objects can be analysed to identify the *sentiment* of the content, and assign a corresponding label, *i.e.* either of *positive*, *neutral* and *negative*.

For the purposes of this research, the back-end was implemented as a separate service running locally on the test computer. It is expected that in a real deployment the back-end functionality will be embedded in the extension itself, both for practical purposes (*e.g.*, offline use) and also to safeguard the user's privacy by ensuring that the analysed text does not leave the host computer.

3.1 Sentiment Analysis

Sentiment Analysis is one of the most widely used applications of ML. A number of good-performing algorithms exist, so much that even Github Copilot's [23] landing page[1] uses *Sentiment Analysis* for its main example.

[1] https://github.com/features/copilot, accessed 22 Dec. 2023

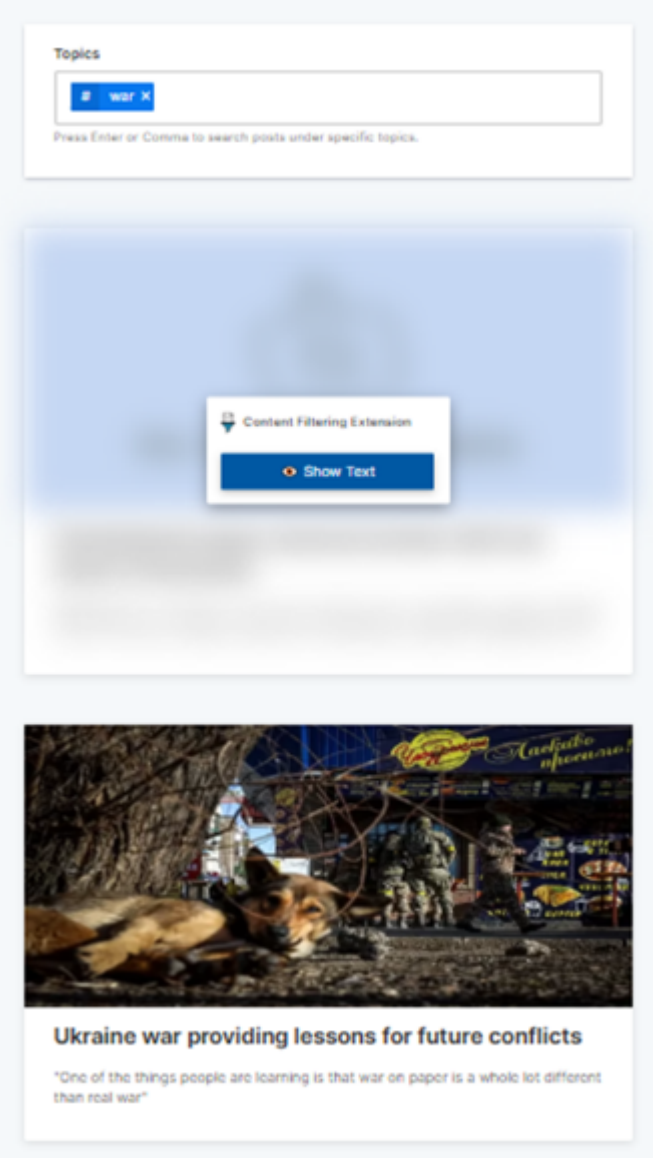

Figure 1: The front-end is responsible for adding a blur mask to items which are identified as *negative*. Non-blocked articles from Euronews (https://www.euronews.com).

The focus of this work is in assessing whether automatically filtered content can improve the well-being of Web users, by limiting access to negative information. As such, a core element of this approach is the use of proper Sentiment Analysis datasets and algorithms. However, the development of such algorithms is beyond the scope of this paper, which instead focuses on identifying and reusing the most suitable datasets and algorithms.

In this section we briefly describe our process for selecting an appropriate dataset for training the algorithm, for preparing by means of data cleaning, and finally the selection of the most effective algorithm to be used in the development of our Web browser extension.

3.1.1 Training Datasets and Data Cleaning. Next, we describe the 3 most suitable datasets for this project and explain why the Sentiment140 dataset was selected to train the ML models.

- *Sentiment140 Dataset* This dataset contains 1.6 million tweets extracted using the Twitter API [13], and it is ranked as one of the most popular datasets for sentiment analysis. All of the tweets were annotated as *positive* or *negative* by utilizing distant supervision, thus avoiding human bias in the sentiment classification. In addition, 350 notebooks were created using this dataset, showing that the community has invested a lot of time analysing and evaluating the dataset. A disadvantage of this dataset is that it does not include neutral data.
- *Twitter Sentiment Analysis Dataset* This dataset[2] contains 32.000 tweets that are labelled as either *hate speech* or *non hate speech*. In this dataset, a tweet is labelled as *hate speech* if it contains racist or sexist language. All the entries were manually labelled by humans. While this is a very useful dataset, especially for its human-powered training, it has an important limitation: it is relatively small, and it has a very specific scope which limits its suitability for our project's objective, *i.e.* detecting overall negativity in text and not only in the specific domain of hate speech. It could however be used as part of a more general dataset to enhance the overall effectiveness.
- *Twitter and Reddit Sentimental analysis Dataset* With a total of 200.000 posts from *Twitter* and *Reddit*, this dataset [6] provides a good match for the needs of this project. All of the posts were labelled as *positive*, *neutral*, or *negative* using pre-trained ML models. However, a limitation of this dataset is that it focuses on people's opinion towards government elections. The specificity of the dataset would affect the effectiveness of the created ML model which is desired to work on the full spectrum of social media posts.

The selected *Sentiment140 Dataset* was then cleaned and prepared for use in the evaluation of the extension. For this purpose, we used best practices from the literature, as described in [19, 25]. Specifically, we used the following techniques to clean and normalise the text input and to assist the overall training and production phases:

- *Replacing URLs*: Links starting with "http", "https" or "www" are replaced by the word "URL".
- *Replacing Emojis*: Emojis are replaced by using a pre-defined dictionary provided by Kolasani and Assaf [18] which lists emojis along with their meaning. For example, ":-)" becomes "EMOJIsmile".
- *Replacing Usernames*: For example, replace "@Username" with the word "USER".
- *Removing Non-Alphabetic Characters*: Replacing non-alphabetic characters with a space.

[2]https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis, last accessed 22 Dec. 2023

- *Removing Consecutive Letters*: Instances of 3 or more consecutive letters are replaced by 2 letters. For example, "Helloooo" becomes "Helloo".
- *Removing Short Words*: Words with length less than 2 letters are removed.
- *Lemmatizing*: Lemmatization is the process of converting a word to its base form. For example, "Running" becomes "Run", and "car's" becomes "car", etc.
- *Lower Case*: All text is converted to lowercase.

3.1.2 Sentiment Analysis Algorithms. For the purposes of the Web browser extension development, we assessed four well-known algorithms for sentiment analysis: Bernoulli Naive Bayes [26, 32], Linear Support Vector Machine (Linear SVC) [35], Logistic Regression [27], and Multinomial Naive Bayes [11, 30].

- For training a ML model using the *Bernoulli Naïve Bayes* algorithm, we used the `BernoulliNB` classifier from *sklearn*. The hyper-parameter `alpha` was set to the optimal value 2, which was determined through a grid search. When the model training was complete, the reusable function `evaluate_model` was called to print a classification report and a confusion matrix regarding the model.
  The classification report of this model indicates an accuracy of 78% when predicting the text polarity of the given training dataset. In addition, based on the f1-score, we can see that the model was effective for predicting both positive and negative sentences.
- For realising the *Linear SVC* algorithm, we used the `LinearSVC` classifier, again from *sklearn*. After adjusting the classifier hyper-parameters to increase the model's accuracy, we found that the default values had the highest accuracy. When the model training was complete, the `evaluate_model` reusable function was called to print a classification report and a confusion matrix regarding the model.
  Like the *Bernoulli Naïve Bayes* classification model, this model has also achieved 78% accuracy when predicting the text polarity of the given training dataset. The f1-score indicates a good prediction level for both positive and negative sentences.
- For training a ML model using the *Logistic Regression* algorithm, we used the `LogisticRegression` classifier, also from sklearn. For this classifier, we had to adjust three of the hyper-parameters in order to achieve the highest accuracy. The hyper-parameter `C` which is the inverse of regularization strength, was set to 2, meaning that the model was instructed to assign a high weight to the training data and a lower weight to the complexity penalty. The hyper-parameter `max_iter` was set to 1.000 since it was the best fit for the model after multiple tries with different values.
  The Logistic Regression model has achieved slightly higher prediction accuracy at 79% compared to the previous two models. Moreover, the f1-score indicates a good prediction level for both positive and negative sentences, similar to the previous models.
- Lastly, for training a ML model using the *Multinomial Naïve Bayes* algorithm, we used the `MultinomialNB` classifier from *sklearn*. After adjusting the classifier hyper-parameter `alpha`, we found that the default value of 1 had the highest accuracy. When the training was complete, the reusable function `evaluate_model` was used to produce the confusion matrix. Lastly, the *Multinomial Naïve Bayes* model has achieved an overall score of 78% of all of the scoring properties of the classification report, indicating that minor adjustments to the hyper-parameters of the classification algorithm could possibly increase the accuracy of the model.

For each of these four algorithms, we assessed them on the selected dataset and produced the *confusion matrices* depicted in figure 2.

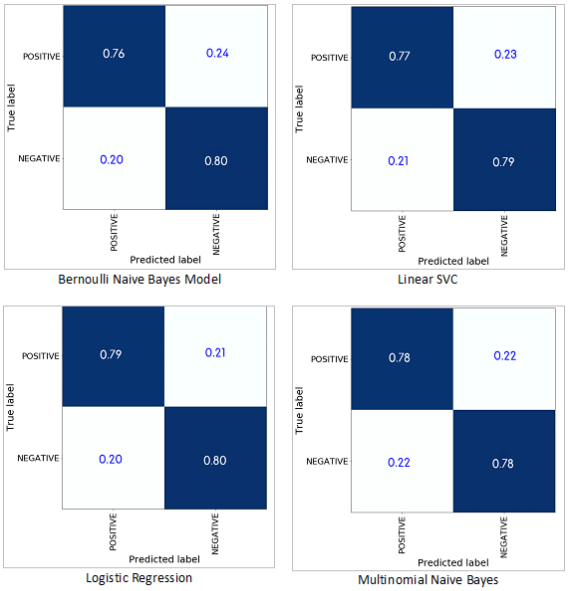

Figure 2: The confusion matrices for the four assessed algorithms.

The evaluation phase showed that all of the models have very similar prediction accuracy. Based on the training data fitted to the models, the most accurate model for classifying the text polarity is the model trained by the *Logistic Regression* classification algorithm with a 79% prediction accuracy, just 1 point more accurate than the other models. In addition, all of the models had an average of 20% *false positives* and *false negatives*, which can be further improved by adjusting the hyper-parameters of the classification algorithms. Finally, the training time of the ML models was insignificant, ranging from 521 to 1506 seconds. Given this, we did not consider this factor in the selection of the most suitable model to be used in this project.

4 EVALUATION

To evaluate the applicability of our approach in real-world use cases, we organised two data collection sessions with a total of 50 participants. In the first session, the participants used the demo website until the extension classified 3 of the viewed articles as negative. Next, the participants were given a survey with questions

related to the sentiment classification accuracy, performance, and its applicability in their personalised everyday lives.

In this section, we discuss the data collection results and apply descriptive analysis to the most relevant variables of the collected data. Additionally, based on similar projects, alternative methods of data collection are evaluated, and improvements in data collection processes and analysis are identified.

Based on the collected data we aim to address the following:

- Does user age affect the acceptance of the extension?
- Does user education affect the acceptance of the extension?
- Do users who assess themselves as noticing mood changes when reading negative news, agree more with the sentiment classification decision of the extension compared to the rest?

4.1 Data Collection

This project's audience consists of English-speaking adults of any gender and background who have some experience with using the Internet via a laptop or desktop computer.

The participants were selected using a non-probability sampling method, in particular convenience sampling, which is an uncomplicated and economical sampling technique. Specifically, the participants were selected based on how easily they could be accessed by the researchers. Admittedly, convenience sampling has its disadvantages: It may produce biased data since the participants are close relatives or friends, resulting in the sampling data being less diverse and thus less representative of the general population. Nevertheless, it is argued that this is suitable for a quick assessment of the proposed approach.

Demographically, the 50 participants chosen for the evaluation of this project were selected from the researchers' workplace, friends, and family circles, and consisted of 28 males and 22 females, ages 19 to 73. All of the participants were briefed about the intention of the research and were asked to answer as honestly as possible.

4.1.1 One-on-one Sessions. One-on-one sessions were conducted in two different locations, and participants were required to use the demo website for as long as it was required until the extension classified three of the viewed articles as negative.

The first location was at a workplace, where 11 male and 10 female co-workers participated in the data collection process. The second location was at a home environment, where 17 male and 12 female members joined the one-on-one sessions and completed the survey.

All of the participants had to select at least one topic or category of news, in order for the demo website to fetch the related articles. As soon as the extension marked three of the articles as negative, the participant were asked to un-blur the articles and read their titles. After reading them, they were prompted to answer the survey.

The same software and hardware was used for all one-on-one sessions. This ensured that the user experience for all participants was as similar as possible. Particularly, a high performance laptop computer was used, with a high speed Internet connection, and the websites were accessed via the Google Chrome web browser.

4.1.2 Survey. After the user testing of the developed extension, we used a survey to collect the user feedback. The survey consisted of 18 questions, both qualitative and quantitative.

Technically the survey was implemented via Google Forms. The participants were able to interrupt and abandon the survey at any time, and were only allowed to complete it once.

Furthermore, to control the quality of the survey and to ensure that every participant was paying attention and not providing random/mindless answers, an attention check question was used. This question was answered correctly by all 50 participants.

In addition, a group of 10 people who did not participate to the actual data collection process, had pretested the survey to ensure that the flow of the survey was logical, the questions and answers were accurate, and there were none or minimal typos and grammar mistakes.

The participants were allowed to answer the survey using their own laptop or mobile phone or the laptop computer used for the tests, but they were asked to do so immediately after they had completed the session in order to answer the questions as accurately as possible.

4.2 Collected Data

In total, we collected 50 entries, one for each participant. This section analyses the collected data, both in terms of demographics, and second in terms of answering the identified research questions.

4.2.1 Demographic Analysis. In terms of *gender* 22 participants responded as "Male", 28 participants as "Female". No participant responded with "Other".

All participants were adults, with 11 of them in the *age group* "18-25", half of them (25) in the group "26-35", and the rest in the groups "36-45" (8), "46-55" (4) and "56+" (2).

Finally, in terms of *education level*, the majority of the participants had a higher education degree: "Bachelor's" (16) or "Master's" (10). For the rest of the participants, their education was at the minimum "high school" (13), and the rest had selected "trade, technical, vocational training" (11).

4.2.2 Assessing the need for the extension. In order to assess the usefulness of the extension assuming it was ideally implemented, we asked the participants to reflect on their personal view of the impact negative news have on them as well as the need for an extension to limit their exposure to such news.

First, the participants were asked to assess whether they "[...] notice any mood changes when reading bad/negative news?". And following this question, the participants were also asked whether they "[...] consider that it would be useful if a tool could filter out unwanted online items?".

The collected results have shown that most participants agreed they are affected by negative news (only 20% stated they do not). Many of them (28%) would consider using a web browser extension that controls and limits the showing of negative news and a majority (56%) would consider it but were not certain. These results are illustrated in figure 3.

4.2.3 Agreement Frequency. An important question challenged the participants to identify "[with] how many of the posts that were flagged by the extension as negative do you agree with?". Given that the one-on-one session was displaying news until exactly 3 news were flagged as negative, the possible answers were 0, 1, 2, and 3.

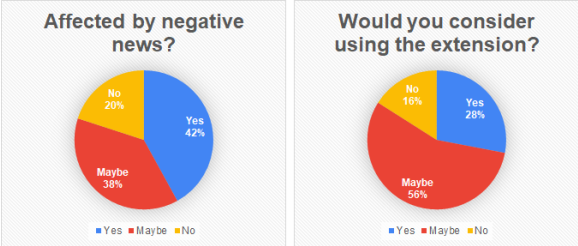

Figure 3: The tendency of the participants: The majority of them have stated they are affected by negative news. Also, more participants would consider the use of such extension than the opposite, with the majority of them being undecided though.

The collected data shows that 28 out of the 50 participants agreed with all of the flagged posts ("3 of 3"), while the remaining participants agreed with some of the flagged posts (specifically 15 agreed with "2 of 3" or 66% and 7 participants agreed with "1 of 3" or 33%). No participant stated that they disagreed with all flagged items ("0 of 3"). These results are also illustrated in Figure 4.

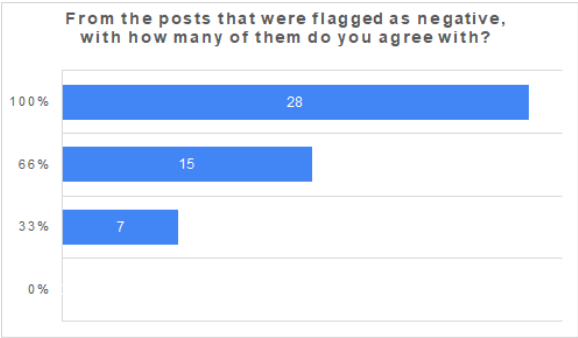

Figure 4: The agreement frequency of the participants: Of the 50 participants, 28 agreed with all "flagged" items, and 22 agreed with some of the flagged items. No participant has disagreed with all flagged items.

4.3 Assessing the User Acceptance of the Proposed Extension

An important question concerns whether the participants have identified a *need* for the proposed extension, and whether they would be selecting to use it if it were available. Specifically, the participants were asked "What is the likelihood of using such a browser extension for your everyday web browsing?".

The results showed that the majority of the participants would use the proposed extension "Very Likely" (12 answers) or "Likely" (25 answers). On the contrary, 13 participants stated it was "Unlikely" they would use this extension, and no one selected "Very Unlikely" as the answer.

Given the demographic data, we have also worked to identify any possible patterns in the collected data, and their impact on this core question. The following subsections present our findings in relation to the four research questions identified at the beginning of this section.

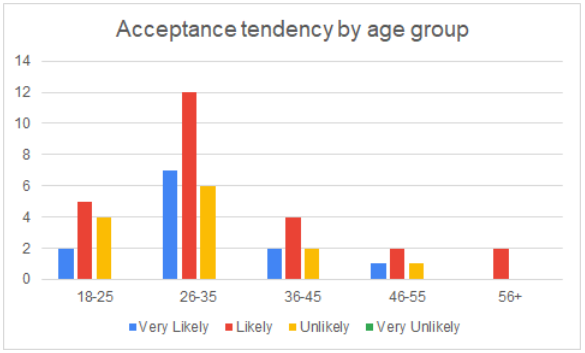

Figure 5: The acceptability of the extension appears to not be significantly impacted by the *age group*: it is 64% in the "18-25" age group, and 75% in the other age groups (except for "56+" which anyway has a very small population of just 2).

4.3.1 Does user age affect the acceptance of the extension? In the first instance we aim to assess whether the *age group* has an impact on the tendency of a participant to use the proposed web browser extension. The results, summarised in Figure 5 show that the age group "26-35" has the highest tendency to accept the use of the extension. The responses of the remaining groups appear to be uniformly distributed, except for the group "56+" which only contains 2 participants.

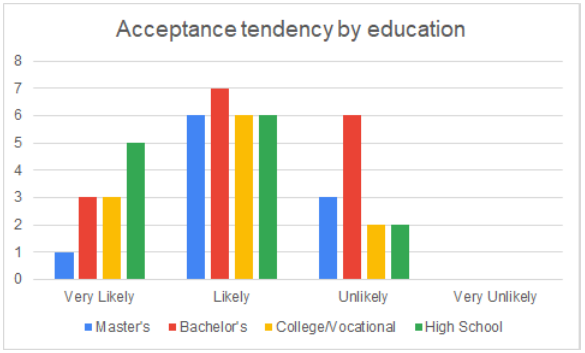

Figure 6: The acceptability of the extension is not significantly impacted by the *education level*: it ranges from 62-70% in participants with a university degree, and 81-83% in participants with a "High School" or "College/Vocational" training degree.

4.3.2 Does user education affect the acceptance of the extension? The education level of the participants was also examined, and it was found that it had some limited impact on the participant's acceptability of the extension. Specifically, it was found that participants with a university degree (*i.e.*, in the "Bachelorś" and "Masterś" groups) had a lower tendency (of 62% and 70% respectively) to adopt the use of the proposed extension, compared to participants

with a "High School" or "College/Vocational" degree (85% and 82% respectively). This is depicted in Figure 6.

4.3.3 Do people who notice mood changes when reading negative news agree more with the sentiment classification decision of the extension? When asked whether they would use the proposed extension, the participants' answer appears to be impacted by whether they feel negative news affect their mood. Specifically, participants who answered that their mood is affected by negative news are less likely to use the extension with 62% (13 out of 21 who answered "Yes") and 79% (15 out of 19 who answered "Maybe") acceptance respectively. The 10 participants who stated their mood is not impacted by negative news said they are more likely to use the proposed extension with an acceptance rate of 90% (9 out of 10). This is shown in Figure 7.

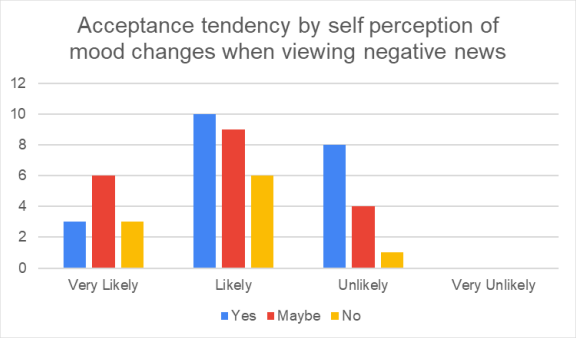

Figure 7: Cross referencing data shows that the acceptance of the extension is affected by whether the participant has answered that their mood is affected by negative news on the Web.

5 DISCUSSION

Our evaluation has revealed that the use of automated Web filtering is indeed needed in the real world. While our evaluation was relatively small, it has nevertheless highlighted some interesting findings:

- The majority of participants accept that they are affected by negative news. This is in line with what we know already about the impact of graphic images from mass media on the mood and well-being of the viewers [1, 3, 4, 7, 24, 31].
- Some participants would use the proposed extension in their everyday web browsing, but the majority were either hesitant or negative. To some extent this is in line with people wishing to be in full control of their activity: This is a well-known fact, and some ML researchers have proposed for "*end-user facing component to provide not only the classification outcome, but also exposing some of the logic of this classification*" [5].
- The majority of the participants also found that the proposed extension was highly accurate. This however was specific to whether the filtered items were negative or not, and not whether they were unwanted or not. Personalisation is key for a widely successful extension of this form. However, this comes with some challenges as well, as collecting data for personalisation may cause some users to perceive it as a privacy violation. For instance, Lee and Cranage have argued that "*behavioural responses are positively related to the perceived usefulness of services and negatively associated with privacy concerns*" [20].
- In terms of "age group" and "education level", it was found that they have no major impact on the participant's acceptance of the extension (*i.e.*, their willingness to use this tool).
- Contrary to our expectation, there was a higher acceptance rate for users who perceived themselves as less affected by negative news. This is an unexpected result, as we would assume that people who feel their mood is negatively affected by some types of content would want to have less exposure to such content. An improvement would be to consider the *mood strait* besides the momentary *mood*. As discussed by Eid, Schneider and Schwenkmezger in [10], "*[whereas] an emotional state characterizes the momentary feeling of an individual, an emotional trait can be defined as the propensity to experience a specific mood or emotion*".

Lastly, a publicly available tool for web filtering would require some additional reflection in terms of its ethical implications. Because of the critical decisions which are increasingly delegated to AI systems, the concern of ethical AI gains immense significance. A survey of AI ethics guidelines by Jobin, Ienca and Vayena [16] has revealed "*a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy)*".

A major ethical issue regarding the training and use of a ML model, or any AI-based prediction algorithm in general, is how was the dataset used to train the generated model. Individual biases on a subject can greatly affect the system's integrity. Additionally, one could ask whether it is ethically correct for an algorithm to decide whether some web-based media content should be hidden or not.

6 CONCLUSIONS

In this paper we presented and assessed an intelligent web filtering tool, in the form of a web browser extension (*i.e.*, plugin) which aims to improve the users' well-being by filtering out negative news. A prototype extension was designed and implemented using standard sentiment analysis algorithms and training data. The developed tool was then evaluated with a user test followed by a survey, involving 50 participants.

The collected data has shown that such a tool is useful and would be welcomed by users. At the same time, the effectiveness of the prototype was assessed and it was found that most participants have rated the automated filtering decisions as correct. Lastly, we assessed whether the acceptance of this extension is affected by demographic or other data, and it was found that it is not affected by age group or education level. At the same time, people who do not think their mood is impacted by negative news have provided a higher acceptance rate for the extension, compared to those who believe their mood is affected.

In the future, we would like to improve the extension by adding the option for the user to learn *why* an item was blurred (*e.g.*, "negative news", "foul language", etc.) and also to allow them to

indicate any *false positives* and *false negatives* so the algorithm can continuously improve and provide more user-tailored results.

REFERENCES

[1] Jennifer Ahern, Sandro Galea, Heidi Resnick, Dean Kilpatrick, Michael Bucuvalas, Joel Gold, and David Vlahov. 2002. Television images and psychological symptoms after the September 11 terrorist attacks. *Psychiatry* (2002), 289–300. Issue 4. https://doi.org/10.1521/psyc.65.4.289.20240

[2] Mark Z. Barabak. [n. d.]. 'Quarantini.' 'Doomscrolling.' Here's how the coronavirus is changing the way we talk. *The Los Angeles Times* ([n. d.]). https://www.latimes.com/world-nation/story/2020-04-11/coronavirus-covid19-pandemic-changes-how-we-talk

[3] Mark Boukes and Rens Vliegenthart. 2017. News Consumption and Its Unpleasant Side Effect. *Journal of Media Psychology* (2017), 137–147. Issue 3. https://doi.org/10.1027/1864-1105/a000224

[4] Luca Braghieri, Ro'ee Levy, and Alexey Makarin. 2022. Social Media and Mental Health. *American Economic Review* 112, 11 (November 2022), 3660–93. https://doi.org/10.1257/aer.20211218

[5] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512. https://doi.org/10.1177/2053951715622512

[6] Kumar A. Chaithanya. 2020. *Twitter and Reddit Sentimental analysis Dataset.* https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset

[7] Sarah M. Coyne, Adam A. Rogers, Jessica D. Zurcher, Laura Stockdale, and McCall Booth. 2020. Does time spent using social media impact mental health?: An eight year longitudinal study. *Computers in Human Behavior* 104 (2020), 106160. https://doi.org/10.1016/j.chb.2019.106160

[8] Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C. Desmarais, Zhen Ming, and Jiang. 2022. GitHub Copilot AI pair programmer: Asset or Liability? https://doi.org/10.48550/ARXIV.2206.15331

[9] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society. https://doi.org/10.14722/ndss.2019.23378

[10] Michael Eid, Christoph Schneider, and Peter Schwenkmezger. 1999. Do you feel better or worse? The validity of perceived deviations of mood states from mood traits. *European Journal of Personality* 13, 4 (1999), 283–306. https://doi.org/10.1002/(SICI)1099-0984(199907/08)13:4<283::AID-PER341>3.0.CO;2-0

[11] Arif Abdurrahman Farisi, Yuliant Sibaroni, and Said Al Faraby. 2019. Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. *Journal of Physics: Conference Series* 1192, 1 (mar 2019), 012024. https://doi.org/10.1088/1742-6596/1192/1/012024

[12] Ana Fonseca and Jorge Osma. 2021. Using Information and Communication Technologies (ICT) for Mental Health Prevention and Treatment. *International Journal of Environmental Research and Public Health* (2021), 461. Issue 2. https://doi.org/10.3390/ijerph18020461

[13] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1, 12 (2009), 2009.

[14] Abid Haleem, Mohd Javaid, and Ravi Pratap Singh. 2022. An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 2, 4 (2022), 100089. https://doi.org/10.1016/j.tbench.2023.100089

[15] Saki Imai. 2022. Is GitHub Copilot a Substitute for Human Pair-Programming? An Empirical Study. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings* (Pittsburgh, Pennsylvania) *(ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 319–321. https://doi.org/10.1145/3510454.3522684

[16] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* (2019). Issue 9. https://doi.org/10.1038/s42256-019-0088-2

[17] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* 99 (2023), 101861. https://doi.org/10.1016/j.inffus.2023.101861

[18] Sai Vikram Kolasani and Rida Assaf. 2020. Predicting stock movement using sentiment analysis of Twitter feed with neural networks. *Journal of Data Analysis and Information Processing* 8, 4 (2020), 309–319.

[19] Muriel Kosaka. 2020. *Cleaning & Preprocessing Text Data for Sentiment Analysis.* https://towardsdatascience.com/cleaning-preprocessing-text-data-for-sentiment-analysis-382a41f150d6

[20] Chung Hun Lee and David A. Cranage. 2011. Personalisation–privacy paradox: The effects of personalisation and privacy assurance on customer responses to travel Web sites. *Tourism Management* 32, 5 (2011), 987–994. https://doi.org/10.1016/j.tourman.2010.08.011

[21] Noble Saji Mathews and Sridhar Chimalakonda. 2021. Detox Browser – Towards Filtering Sensitive Content On the Web. https://doi.org/10.48550/ARXIV.2106.09937

[22] Seong Jae Min. 2019. From algorithmic disengagement to algorithmic activism: Charting social media users' responses to news filtering algorithms. *Telematics and Informatics* 43 (2019), 101251. https://doi.org/10.1016/j.tele.2019.101251

[23] Nhan Nguyen and Sarah Nadi. 2022. An Empirical Evaluation of GitHub Copilot's Code Suggestions. In *Proceedings of the 19th International Conference on Mining Software Repositories* (Pittsburgh, Pennsylvania) *(MSR '22)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3524842.3528470

[24] Ayokunle A. Olagoke, Olakanmi O. Olagoke, and Ashley M. Hughes. 2021. Exposure to coronavirus news on mainstream media: The role of risk perceptions and depression. *British journal of health psychology* (2021), 865–874. Issue 4. https://doi.org/10.1111/bjhp.12427

[25] Saurav Pradha, Malka N. Halgamuge, and Nguyen Tran Quoc Vinh. 2019. Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*. 1–8. https://doi.org/10.1109/KSE.2019.8919368

[26] Atiqur Rahman and Md. Sharif Hossen. 2019. Sentiment Analysis on Movie Review Data Using Machine Learning Approach. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*. 1–4. https://doi.org/10.1109/ICBSLP47725.2019.201470

[27] W.P. Ramadhan, S.T.M.T. Astri Novianty, and S.T.M.T. Casi Setianingsih. 2017. Sentiment analysis using multinomial logistic regression. In *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*. 46–49. https://doi.org/10.1109/ICCEREC.2017.8226700

[28] Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner. 2019. Survey and Benchmarking of Machine Learning Accelerators. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*. 1–9. https://doi.org/10.1109/HPEC.2019.8916327

[29] Albert Reuther, Peter Michaleas, Michael Jones, Siddharth Samsi, Jeremy Kepner, and Vijay Gadepally. 2020. Survey of Machine Learning Accelerators. In *2020 IEEE High Performance Extreme Computing Conference (HPEC)*. 1–12. https://doi.org/10.1109/HPEC43674.2020.9286149

[30] Gurinder Singh, Bhawna Kumar, Loveleen Gaur, and Akriti Tyagi. 2019. Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*. 593–596. https://doi.org/10.1109/ICACTM.2019.8776800

[31] Zahava Solomon, Karni Ginzburg, Avi Ohry, and Mario Mikulincer. 2021. Overwhelmed by the news: A longitudinal study of prior trauma, posttraumatic stress disorder trajectories, and news watching during the COVID-19 pandemic. *Social science & medicine* (2021). Issue 113956. https://doi.org/10.1016/j.socscimed.2021.113956

[32] Junseok Song, Kyung Tae Kim, Byungjun Lee, Sangyoung Kim, and Hee Yong Youn and. 2017. A novel classification approach based on Naïve Bayes for Twitter sentiment analysis. *KSII Transactions on Internet and Information Systems* 11, 6 (June 2017), 2996–3011. https://doi.org/10.3837/tiis.2017.06.011

[33] Bernd Carsten Stahl and Damian Eke. 2024. The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management* 74 (2024), 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

[34] Nurul Athirah Binti Suliman and Hazinah Binti Kutty Mammi. 2017. Explicit words filtering mechanism on web browser for kids. In *2017 6th ICT International Student Project Conference (ICT-ISPC)*. 1–6. https://doi.org/10.1109/ICT-ISPC.2017.8075322

[35] Nikhil Yadav, Omkar Kudale, Aditi Rao, Srishti Gupta, and Ajitkumar Shitole. 2021. Twitter Sentiment Analysis Using Supervised Machine Learning. In *Intelligent Data Communication Technologies and Internet of Things*, Jude Hemanth, Robert Bestak, and Joy Iong-Zong Chen (Eds.). Springer Singapore, Singapore, 631–642.

[36] Burak Yetistiren, Isik Ozsoy, and Eray Tuzun. 2022. Assessing the Quality of GitHub Copilot's Code Generation. In *Proceedings of the 18th International Conference on Predictive Models and Data Analytics in Software Engineering* (Singapore, Singapore) *(PROMISE 2022)*. Association for Computing Machinery, New York, NY, USA, 62–71. https://doi.org/10.1145/3558489.3559072

[37] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* 8 (2020), 58443–58469. https://doi.org/10.1109/ACCESS.2020.2983149

Received 2023; revised 12 March 2024; accepted 5 June 2024