Incremental Learning for Fake News Detection

Renato M. Silva[1], Pedro R. Pires[2], Tiago A. Almeida[2]

[1] Department of Computer Engineering, Facens University, Sorocaba, São Paulo, Brazil
renato.silva@facens.br
[2] Department of Computer Science, Federal University of São Carlos (UFSCar), Sorocaba, São Paulo, Brazil
pedro.pires@dcomp.sor.ufscar.br, talmeida@ufscar.br

Abstract. Fake news is a concern that has impacted people's lives for a long time. However, this problem has worsened deeply with the increase of social media popularity, which became a fertile ground to spread fast and affect humanity's social, political, and economic future. Despite several studies on fake news detection, some critical gaps still need to be addressed. One of them is that most studies are unrealistic since they use machine learning with offline learning models. The language used in communication change continuously, reflecting society's nature. Therefore, as facts covered by the news are dynamic, the static models learned by offline learning methods can quickly become obsolete. This study evaluates fake news detection using the online learning paradigm, which is best suited for dynamic problems whose underlying data distribution can change over time. We have addressed how automatic fake news classification suffers from concept drifting. For this, we have applied state-of-the-art methods that can learn incrementally to classify documents covering two historical events: the United States presidential election and the coronavirus disease (Covid-19) pandemic. We also evaluated three different types of feedback (uncertain, delayed, and immediate) and two training strategies: (*i*) updating the model only when it makes a prediction error and (*ii*) updating it after both error or success. The results obtained by our carefully designed experiments indicated that the performance of online learning models improved over time, while offline models did not sustain their performance.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: fake news, online learning, text categorization, machine learning

1. INTRODUCTION

Fake news is an ever-growing problem of our time. The popularization of social media and instant messaging apps made the spread of deceptive articles more accessible, cheaper, and faster. As a result, humans are exposed daily to fake news [van der Linden et al. 2020]. This sudden increase in the volume and impact of fake news is problematic since it can spread misinformation, affect people's opinions and political choices, damage the reputation of public figures, and even incite violence [Zhou and Zafarani 2020]. However, tackling the problem is not easy, mainly because they spread fast uncontrollably. This is due to technological factors, the way people consume information, manipulation, truth bias, and because a person is more likely to share falsity than truth [Charles F. Bond and DePaulo 2006; Vosoughi et al. 2018].

Exposure to fake news tends to increase during significant events, such as a presidential election or a global pandemic, which can cause harm to democracy and even endanger public health [Allcott and Gentzkowf 2016; Zhou and Zafarani 2020]. For instance, an empirical study conducted by [Galhardi et al. 2020] about the Covid-19 pandemic analyzed user reports sent to the *Eu Fiscalizo* Brazilian application from March 17 to April 20, 2020. The authors noticed that more than half (65%) of the received reports consisted of articles presenting homemade and inaccurate procedures to prevent the

We gratefully acknowledge the support provided by the São Paulo Research Foundation (FAPESP; grant 2018/02146-6).

spread of the virus, while 20% contained methods to cure the disease. The authors then concluded that the spread of fake news could "discredit science and global public health institutions and weaken people's adherence to the necessary preventive care when addressing the epidemic". As a result, the fake news problem became so severe during the pandemic that journalists and experts coined the term "infodemic". Simultaneously, the world faced a pandemic of misinformation and the coronavirus pandemic [Zarocostas 2020; Salvi et al. 2021]. This may have encouraged behaviors that undermined efforts by governments and health authorities to implement preventive measures, increasing the severity of the problem [Galhardi et al. 2020; Salvi et al. 2021].

One of the first steps to prevent the spread of fake news is detecting and filtering misleading information. Machine learning methods are one of the leading studied solutions, with different proposed strategies. Many studies have focused on creating fake news datasets [Monteiro et al. 2018; Wang 2017], evaluating different algorithms [Alves et al. 2019; Silva et al. 2020], and conducting feature engineering over linguistic-based attributes [Zhou et al. 2004; Shu et al. 2017]. However, most of these studies do not consider the dynamic nature of the problem. Some even used datasets without information about when the document was posted [Faustini and Ferreira Covões 2019; Wang 2017; Ghosh and Shah 2018]. Moreover, most of the experiments followed an offline learning paradigm, in which batches of data are fed to the model, neglecting temporal information [Biesialska et al. 2020]. Usually, the methods cannot update their model incrementally (also known as offline learning methods) [Silva et al. 2020; Biesialska et al. 2020; Pérez-Rosas et al. 2018; Zhou et al. 2020]. Nevertheless, in real-world scenarios, this behavior is undesirable. For example, relevant topics constantly shift in news media, and new terms can arise while old ones become outdated [Zhang and Kejriwal 2019]. This may result in underlying data distribution changes, a problem known as the concept drift phenomenon [Horne et al. 2019; Ksieniewicz et al. 2020].

Studies using offline learning ignore the temporal variation of news. As a consequence, the results often are overestimated and unreliable. The proper way to conduct an experiment on a task affected by temporal information, such as fake news detection, is through an online learning paradigm. In online learning, a continuous stream of training examples is provided in sequential order, with feedback given after the inference stage [Biesialska et al. 2020]. Incremental learning methods are the most suited for that kind of problem since their training process can occur even in scenarios of scarce memory, and the predictive model can quickly adapt to changes in the input data through constant updates [Silva et al. 2017].

The training pipeline in the offline learning paradigm, also known as the *prequential approach* or *interleaved test-then-train* [Gama et al. 2013; Faceli et al. 2021], is as follows: first, data is fed individually to the classifier. Then, the model generates a judgment and compares it with the received feedback containing the gold standard class. In the same way, as in the offline learning approach, the classifier can then update the predictive model based on the known class. This process continues, with the classifier receiving more samples to predict and its appropriate feedback, one at a time, improving the model. In most studies using the prequential approach, feedback is given immediately after the prediction stage, which may not correspond to real-world scenarios. Commonly, the feedback occurs moments later in real applications — or may even not occur at all. Therefore, to increase confidence in the obtained results, some studies on online learning problems — especially spam filtering — consider different feedback types. Two common approaches are the use of delayed and uncertain feedback. In the first, the gold standard is given to the classifier sometime later, not necessarily following a pattern, while in the second, the classifier only receives feedback for a portion of the data [Cormack 2007; Bittencourt et al. 2020]. Although the type of feedback can affect the outcome of the classifier [Bittencourt et al. 2020], few studies have investigated its impact on the fake news classification problem.

In a previous study [Silva and Almeida 2021], we evaluated how concept drift can impair the classification of fake news. However, we analyzed only one online learning method: the passive-aggressive (PA). Moreover, we considered only the scenario in which feedback is received when the model predicts the wrong labels. The results obtained indicated that the performance of offline models is over-optimistic since the accuracy of the predictions dropped drastically when the concept drift phenomenon occurred. Alternatively, experiments with PA indicated that online learning models could adapt to changes in textual patterns over time, endorsing our hypothesis.

This paper is an extension of our previous paper [Silva and Almeida 2021] and offers a more robust study on online learning models to address fake news detection. First, we conducted experiments with three other baseline online learning models: multinomial naïve Bayes, perceptron, and stochastic gradient descent. Moreover, we have included the scenario where the feedback is given even if the model predicts the correct label. Finally, we again considered different types of online learning feedback, i.e., immediate, delayed, and uncertain feedback [Silva and Almeida 2021]. With these new analyses, we aim to investigate if the conclusions obtained by other studies that disregarded the dynamic nature of the news and by our previous work, which used only one online learning method, are sustained.

Considering our initial experiments and the new content presented here, we intend to fill the gaps in the literature and answer the following research questions:

(1) How concept drift affects the fake news classification problem?
(2) Is the performance of methods trained with the offline learning paradigm sustained by changes in the news patterns?
(3) How far is the performance of the realistic training on error approach compared with the utopic constant feedback?
(4) How do different types of feedback impact the performance of incremental learning methods applied to the classification of fake news?

The remainder of this paper is organized as follows. The next section presents the main related work in the area. The materials and methods are described in Section 3. Section 4 reports our experiments and the obtained results. Finally, conclusions and guidelines for future work are presented in Section 5.

2. RELATED WORK

In recent years, fake news detection has seen significant advancements mainly due to recent techniques in natural language processing and other related areas, with many strategies on how to tackle the problem being proposed. The two main approaches used to deal with fake news are the linguistic-based or content-based features [Shu et al. 2017]. The first one uses linguistic attributes to represent textual samples, such as grammatical classes, semantics, spelling errors, expressivity, and content diversity. In a pioneer study, [Zhou et al. 2004] proposed several features when depicting deceptive texts for classification tasks. Inspired by their work, [Monteiro et al. 2018] constructed a dataset of fake news in the Portuguese language, using attributes such as pausality, emotiveness, uncertainty, and non-immediacy. On the other hand, the content-based approach employs techniques such as part of speech tags, syntactic information, readability metrics, term frequency, and word semantic classes [Pérez-Rosas and Mihalcea 2014; 2015; Pérez-Rosas et al. 2018]. In this kind of approach, samples are usually represented in a distributive or distributed manner. Studies using distributive representation often apply two bag-of-words (BoW) representation types: term-frequency or term-frequency-inverse document frequency (TF-IDF) [Silva et al. 2020]. The latter generally achieves better results [Salton and Buckley 1988]. On the other hand, studies using distributed representation usually employ state-of-the-art neural embeddings models based on context and word co-occurrence, e.g., Word2Vec [Silva et al. 2020; Song et al. 2021; Wang 2017], FastText [Silva et al. 2020; Alves et al. 2019], and GloVe (global vectors) [Kaliyar et al. 2020].

Traditionally, fake news detection is modeled as a binary classification problem. The goal of the classifier is to predict whether a news article is legitimate or fake [Silva et al. 2020; Monteiro et al.

2018]. Some studies use one-class learning, focusing only on the interest class (*i.e.*, fake news) [Gôlo et al. 2021]. However, there are many cases in the real world where news is neither wholly true nor completely false. A common strategy among news agencies, especially in sensationalist newspapers and blogs, is reporting rumors or half-truths [Rezayi et al. 2018]. This is used mainly in political articles, where one can spread hoaxes and rumors about their most minor favorable candidates so that the content seems genuine [Allcott and Gentzkowf 2016]. To address this problem, some studies consider fake news detection as a multiclass or multilabel classification task. For instance, [Rasool et al. 2019] divided the labels into two levels. In the first level, the document is classified as true or false. In contrast, in the last level, the positive class is classified as "mostly-true", "true", "barely-true", or "half-true", and the negative class as "false" or "pants-fire" (news that presents absurd fake information). A different formulation was also used in the Fake News Challenge, in which articles were classified as "agrees", "disagrees", "discusses", and "unrelated", according to the relation between the headline and the body text [Kaliyar et al. 2019].

Many different learning algorithms have been applied to automatically detecting fake news. Traditional methods, commonly employed in other NLP tasks, have been extensively used in different studies, such as support vector machines [Silva et al. 2020; Rasool et al. 2019; Monteiro et al. 2018; Zhou et al. 2020; Cardoso et al. 2018], naïve Bayes [Almeida et al. 2011; Silva et al. 2020; Zhou et al. 2020; Alberto et al. 2015b; 2015a], random forest [Silva et al. 2020; Zhou et al. 2020; Horne et al. 2019] K-nearest neighbours [Silva et al. 2020; Almeida et al. 2016], and logistic regression [Silva et al. 2020; Zhou et al. 2020]. Recently, neural models are gaining ground in the literature, with complex artificial neural networks achieving promising results, e.g., long short term memory (LSTM) [Alves et al. 2019; Wang 2017] and convolutional neural networks (CNN) [Wang 2017; Kaliyar et al. 2020; Khan et al. 2021]. In addition, advanced language models, such as bidirectional encoder representations from Transformers (BERT), have been studied [Khan et al. 2021]. However, with the problem being constantly more present — and harmful — in our digital lives, opportunities for improvements are still very much desirable.

Although there are many advancements in intelligent algorithms and feature engineering for articles representation, most existing studies do not consider the chronological order of news, evaluating the methods unrealistically through an offline learning paradigm [Silva et al. 2020; Zhou et al. 2020; Rasool et al. 2019; Kaliyar et al. 2020; Wang 2017]. In this approach, the date when the article was published is ignored. The model is trained with a fixed batch of news and evaluated on another batch, without a proper analysis of how temporal dynamics affect the prediction's accuracy. This contradicts one of the primary aspects of news in the real world: its dynamism. As new facts constantly occur, new terms that refer to politicians, celebrities, companies, and technologies arise frequently, and an ever-changing relevance is given to specific topics. With that in mind, we hypothesize that offline approaches are not appropriate for fake news detection in real-world scenarios.

Other studies in the literature support our hypothesis. For example, [Horne et al. 2019] evaluated the impact of time in fake news classification. The authors trained a state-of-the-art random forest model with distant labeling to predict whether an article was trustful or unreliable. Different datasets were used, with samples represented with linguistic-based features. The obtained results showed how the performance of the classifiers slowly degrades as time progresses. The authors tried two different strategies to alleviate the problem, online learning and Dynamic Weighted Majority (DWM) [Kolter and Maloof 2007], with the former being sufficient to diminish the effects of concept drift.

[Zhang and Kejriwal 2019] also conducted a study on temporal influence. The authors analyzed the underlying data distribution changes in two tasks related to fake news detection: bias and sensationalism detection. They trained two learning algorithms– logistic regression and support vector machines– over different datasets of articles extracted in 2017 and 2019 and evaluated them in various scenarios. When using models trained on the 2017 dataset to predict samples of 2019, the obtained results demonstrated how terms can become outdated and affect the outcome and how concept drift is occurring at a faster pace in more recent news.

Lastly, [Ksieniewicz et al. 2020] proposed novel classification methods based on a feature extraction strategy to address fake news detection in streaming data from social media. They evaluated the methods through a prequential methodology (or "test-then-train" methodology, commonly used in online learning): the algorithms are tested over an incoming portion of data, not seen during the training phase, and then updated with the original labels, in an alternate manner. The classifiers they applied in the experiments were Gaussian naive Bayes, multi-layer perceptron, and Hoeffding tree. The results showed that the quality of the classifications stabilized over time when they used online learning strategies.

Although all studies mentioned above evaluate the impact of concept drift in fake news detection, there are still gaps and issues that demand more attention. For example, [Horne et al. 2019] used only linguistic-based features instead of the full content, and the algorithms were selected based on their performance over general problems involving concept drift [Minku et al. 2010]. However, this assumption may not hold in the fake news detection scenario. In addition, [Zhang and Kejriwal 2019] performed experiments using an offline learning paradigm, with classifiers commonly used in offline learning problems (logistic regression and support vector machines). Finally, although the study of [Ksieniewicz et al. 2020] shares similar goals with our, the dataset used, *Getting Real about Fake news*[1], composed of 13,000 articles scraped from 244 websites tagged as "bullshit" by the BS Detector Chrome Extension, only extends on a period between October 25, 2016, and November 25, 2016. We believe that data from just one month may not be adequate to properly analyze the influence of concept drift on fake news detection, requiring confirmation of a study considering a more extended period.

Excluding the experiments of [Ksieniewicz et al. 2020], none of the related work presented in this section used distributive text representation to evaluate concept drift. Additionally, even in studies where more than one learning algorithm was evaluated, there were no comparisons on how different models behave when dealing with concept drift problems. Usually, the analysis is conducted only on the occurrence of concept drift itself. Moreover, to the best of our knowledge, no study analyzed the fake news classification using the online learning paradigm with different types of feedback. In online learning, the feedback given to the model after the inference stage can be captured in distinct ways, and results in other NLP problems show how this decision can influence the results [Bittencourt et al. 2020]. Studying this effect on the fake news problem may reveal opportunities for improvement in the classification quality.

3. MATERIALS AND METHODS

To assess the concept drift in the classification of fake news, we have used a set of news published over a long period. As mentioned in Section 2, most studies use datasets that do not contain temporal information or span a short time. Based on these constraints, we performed experiments with the followings datasets:

—NELA-GT-2019[2] [Gruppi et al. 2020]: it contains 1,200,000 news articles from 260 sources published between January 1st, 2019, and December 31st, 2019.

—NELA-GT-2020[3] [Gruppi et al. 2021]: it contains 1,779,127 news articles from 519 sources published between January 1st, 2020, and December 31st, 2020. This dataset includes a subset of news about Covid-19 and another subset with 2020 US presidential election-related articles.

[1] *Getting Real about Fake news.* Available at `https://www.kaggle.com/mrisdal/fake-news` (accessed on January 13, 2023).
[2] NELA-GT-2019. Available at `https://doi.org/10.7910/DVN/O7FWPO` (accessed on January 13, 2023).
[3] NELA-GT-2020. Available at `https://doi.org/10.7910/DVN/CHMUYZ` (accessed on January 13, 2023).

The news from NELA-GT-2019 and NELA-GT-2020 are labeled as unreliable, mixed, or reliable. We have removed from the class mixed and performed experiments only with the news labeled as unreliable or reliable, as with other studies that address fake news, such as [Ksieniewicz et al. 2020] and [Horne et al. 2019],

In the tokenization process, we converted all documents to lowercase and used non-alphanumeric characters as delimiters (except underscore). Moreover, in all experiments, we used TF-IDF [Salton and Buckley 1988] to create a vector representation of each document.

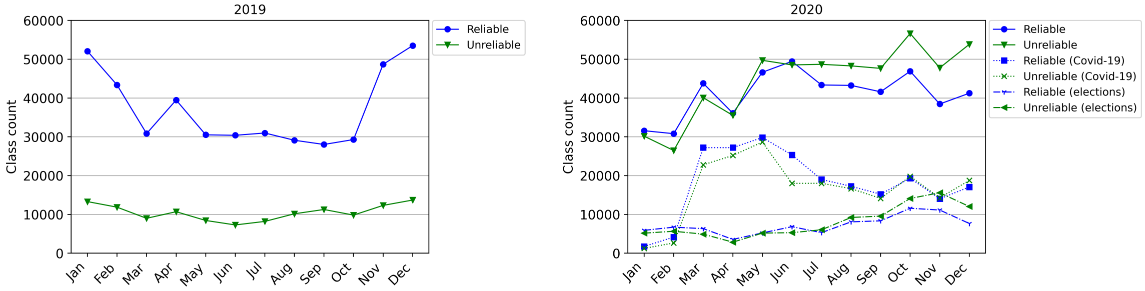


Fig. 1. Number of news from each class.

Figure 1 presents the number of news from each class in each month of 2019 and 2020. The number of unreliable news in 2019 is much smaller than the number of reliable ones. Nevertheless, in almost every month of 2020, the proportion of true news and fake news is similar. Another interesting point is that the number of fake news about the Covid-19 pandemic was higher in the first months of the pandemic. In contrast, the number of fake news about the presidential election increased with the approach of the elections.

The best-known online learning methods were compared: passive-aggressive (PA) [Crammer et al. 2006], multinomial naïve Bayes (M.NB) [McCallum and Nigam 1998], perceptron [Freund and Schapire 1999], and Stochastic Gradient Descent (SGD) [Zhang 2004]. We used the implementation from *scikit-learn library*[4] with the default parameters. To compare the results, we employed the traditional F-measure.

To investigate how concept drift may affect the fake news classifier, we analyzed how non-ordinary or periodic impacting events like the Covid-19 pandemic and the US presidential election impair the performance of methods trained on an offline learning paradigm. We performed experiments with the offline and online learning paradigms to properly answer the research questions, using the protocol presented in Figure 2.

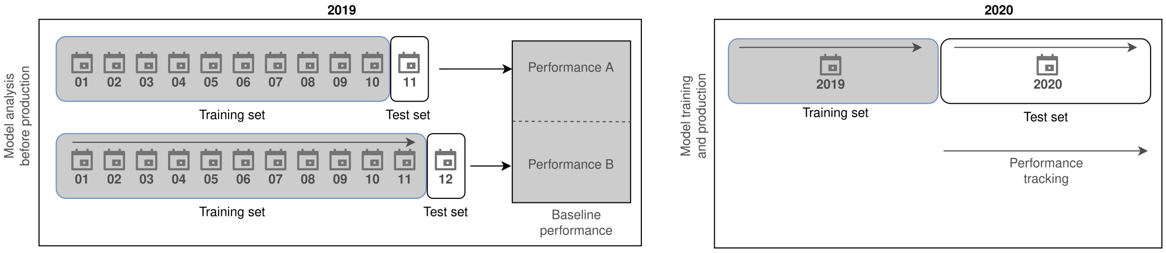


Fig. 2. Experimental protocol (Source: [Silva and Almeida 2021]).

[4] *Scikit-learn.* Available at `http://scikit-learn.org/stable/` (accessed on January 13, 2023).

We first used a method to train a classification model with news from January to the end of October 2019 and test it with the news from November 2019. Next, we used the same method to train another classification model with news from January to the end of November 2019 and test it with the news from December 2019, comparing both performances. If there is no concept drift on these data, we expect that the performance achieved in December will be close to that obtained in November. In this way, it is safe to use this performance as an expected baseline in future data. Finally, we trained a model with all news from 2019 and tested it with the news from 2020, continuously tracking the model performance.

These experiments follow the traditional and real-world classification paradigm, where the model is first statically trained, evaluated, and then put into production. The results obtained with the two test sets, composed of news from November (Performance A) and December 2019 (Performance B), are used as a baseline (expected performance). If the expected performance is kept when classifying the news from January to December 2020, we can safely conclude that the model is not affected by the concept drift.

In the offline paradigm, the model is kept static during 2020. Thus, it is not robust against concept drift. On the other hand, the model is updated continuously in the online paradigm, allowing it to adapt to changes in the data patterns. In this case, the model can be updated with different types of feedback, as explained in the next section.

3.1 Online learning

To investigate how impacting events like the COVID-19 pandemic and the US presidential election affect the performance of methods trained on an online learning paradigm, we designed the experiments as shown in Figure 3.

We used the news from 2019 for training the classifier. In the test stage, we used the news from 2020 through the *prequential approach* [Gama et al. 2013]: we present the documents one at a time to test the classifier, which makes a prediction; then, the classifier can receive feedback and, based on the gold standard class, it can update its predictive model. Although we update the predictive model over time in this scenario, the word dictionary obtained during training was not updated because the implementation of the classification methods we used in this study expects feature vectors with a fixed size.

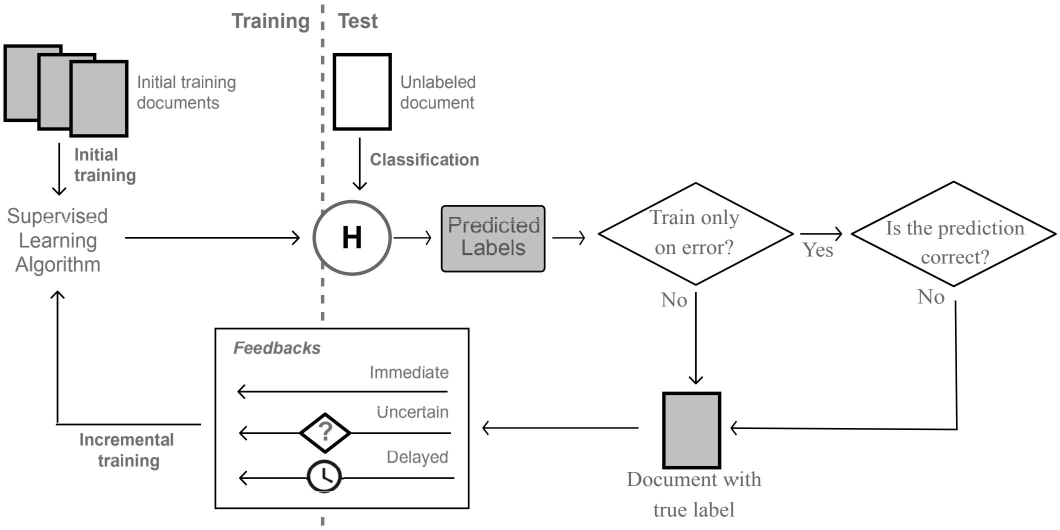


Fig. 3. Overview diagram of the online learning scenario (Source: adapted from [Bittencourt et al. 2019]).

To simulate real-world scenarios, we performed experiments varying the way the feedback is presented, following the protocol defined in previous studies to evaluate online learning approaches [Almeida and Yamakami 2012; Cormack 2007; Bittencourt et al. 2020]:

—**Immediate feedback:** it simulates an ideal scenario, where after receiving and classifying a document, the classifier immediately receives feedback and updates its predictive model;

—**Uncertain feedback:** the classifier receives feedback only for some documents; and

—**Delayed feedback:** the classifier receives feedback with a delay.

We performed experiments considering a scenario where the feedbacks presented above are provided only when the classifier predicts a wrong label (i.e., training on error), and another one where the classifier receives feedback even when the classification is correct. Naturally, the first scenario is much more realistic. However, we used the second scenario as an ideal one which we can consider as a baseline (or target) scenario. Furthermore, in experiments with uncertain feedback, whether to send feedback or not is defined randomly. Finally, in the delayed feedback, the delay is randomly set between 0 and 20 messages.

In a real-world application, the feedback can vary. For example, the frequency of feedback on a fake news filtering application installed on mobile devices of regular users would likely be different from the frequency obtained on an application used by a news agency. We believe the feedback in mobile applications is likely to be similar to that given by email users [Cormack 2007]. Hardly a user would present the behavior simulated by the immediate feedback scenario, immediately correcting a wrong prediction given by the fake news filter. Thus it offers an optimistic overall performance. In a more realistic setting, some users would present the behavior simulated in the delayed feedback scenario, correcting the wrong predictions with a delay. Finally, some users would also present the behavior simulated in the uncertain feedback scenario, correcting the wrong predictions only for some news. In an application used by a news agency, we believe that prediction errors would be corrected more often but would probably follow the scenario simulated by the delayed feedback scenario due to limited human resources.

4. RESULTS

For every scenario described in Section 3, we computed the expected performance (baseline) considering the last two months of 2019 and the F-measure of the classification throughout 2020, presented below. In every figure, the highlighted area with a gray background corresponds to the expected performance (baseline) computed in the last two months of 2019. For example, in the experiments with PA and Perceptron, the expected F-measure for the next months is around 0.7. For M.NB, the expected value is around 0.60, and for SGD is around 0.53. If we naively assumed that the future data has no concept drift, we would expect the performance of the classifiers throughout 2020 will keep similar to the gray region. To answer the research question (1), we must analyze how the results behave during 2020, especially after impacting events such as the Covid-19 pandemic and the US presidential elections.

Figure 4 presents the F-measure obtained in the experiments with the offline learning paradigm. The results aim to answer the research question (2), regarding the robustness of the performance obtained by the classifiers in the offline learning paradigm when they are presented with changes in the news pattern. As we can see, the classifiers' performance suffered a considerable drop in 2020, probably because of a concept drift caused by news about Covid-19. The content of these documents may have different data patterns like medical terminology, drugs, and treatments, not seen in the training set. For example, the performance obtained by PA in January 2020 was 39% lower than in December 2019, and in the month with the worst performance (April 2020), the difference to December 2019 was 47%. In the experiments with SGD, the drop in performance was even more significant: its

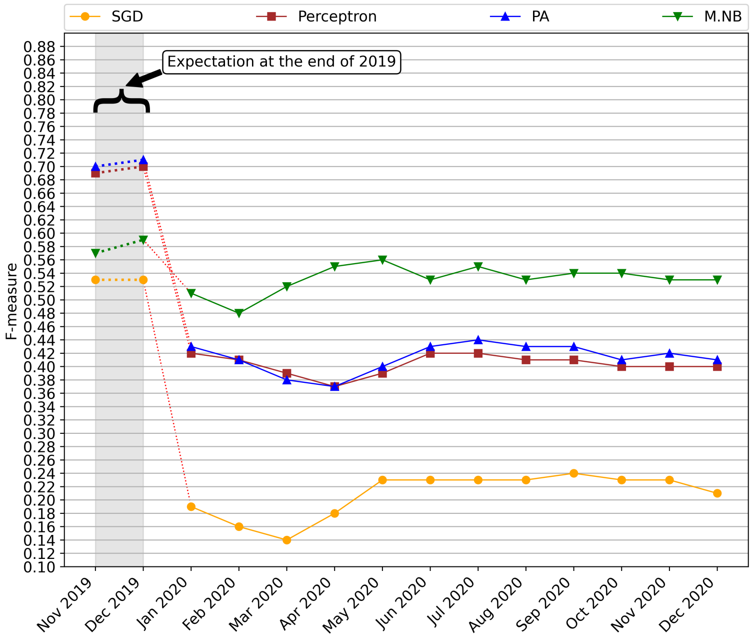


Fig. 4. F-measure obtained in the offline learning scenario. Data is fed in batches and the model is kept static thorough the experiment. Gray area indicates the expected performance considering the last two months of 2019.

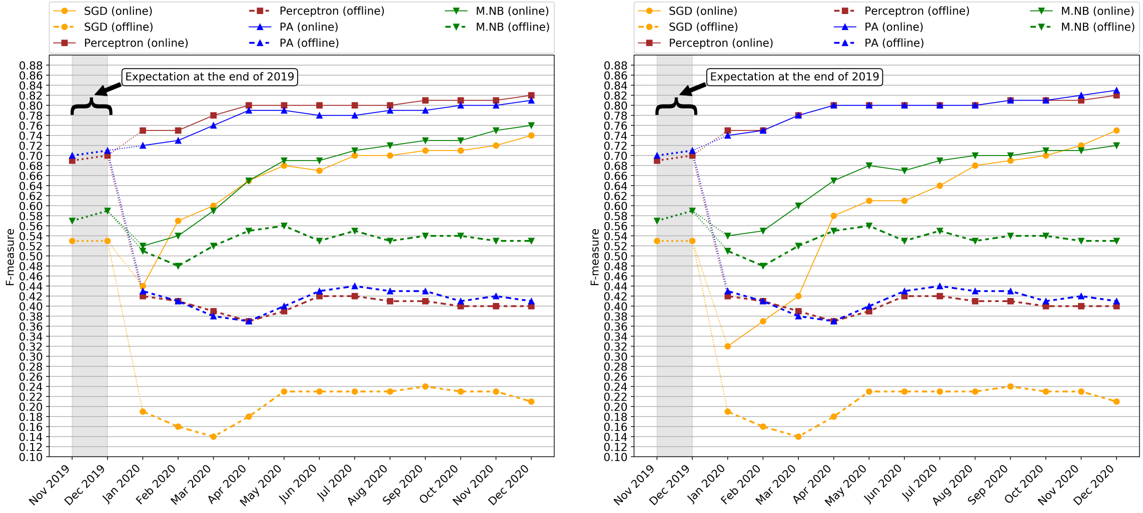
performance in January 2020 was 64% lower than in December 2019. In the month with the worst performance of SGD (March 2020), the difference to December 2019 was 74%. The method with the smallest performance drop in 2020 was M.NB. These results indicate that the expected performance was overestimated at the end of 2019. Therefore, we can safely conclude that a static classification model, which does not adapt to changing news patterns over time, is unsuitable for detecting fake news in real-world scenarios. These results also show that studies based on offline learning models can present overestimated results.

Figures 5, 6 and 7 presents the performance obtained by the classifier in the online learning scenarios considering the three types of feedback: immediate, uncertain, and delayed.

While the performance of the offline learning paradigm presented a significant drop when classifying the 2020 news, the results in the online learning paradigm were initially similar to the results obtained at the end of 2019 and even increased over time. In December 2020, the performance of Perceptron trained with the offline learning paradigm was about 51%, on average, lower than the performance of this method using the online learning paradigm with immediate feedback.

To address the research question (3), we can compare the results obtained in the scenarios where feedback was given only on an error with the one when the model receives feedback on correct predictions. The classifiers obtained similar results in both of them. However, the best performance is obtained in some experiments when the feedback is sent only in error. This result is interesting because it is more realistic to send feedback only when the classifier makes a prediction error in real applications. Furthermore, the results indicated that this process does not lose much performance even though it is straightforward.

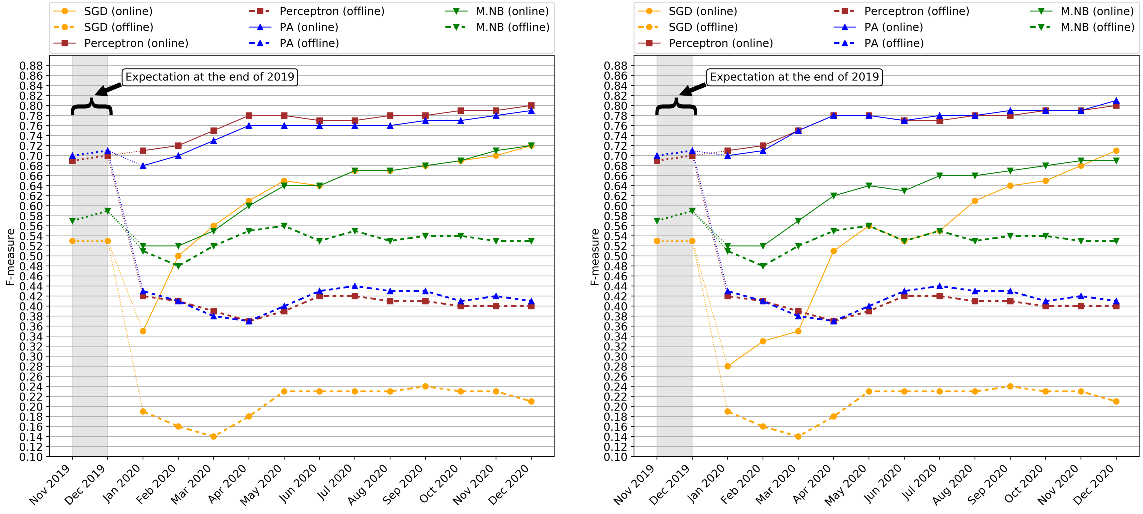
Finally, regarding the research question (4), we can analyze the F-measure obtained for each type of feedback. The best results were obtained when the immediate feedback scenario was used, which was expected since it is the ideal but over-optimistic scenario. The delayed feedback was the one that most negatively affected the performance. It was lower than the one obtained in the immediate



(a) Training only on error.

(b) Training with all the news.

Fig. 5. F-measure obtained in the online learning scenario with immediate feedback. After each prediction, the classifier receives feedback and immediately updates the model. Gray area indicates the expected performance considering the last two months of 2019.

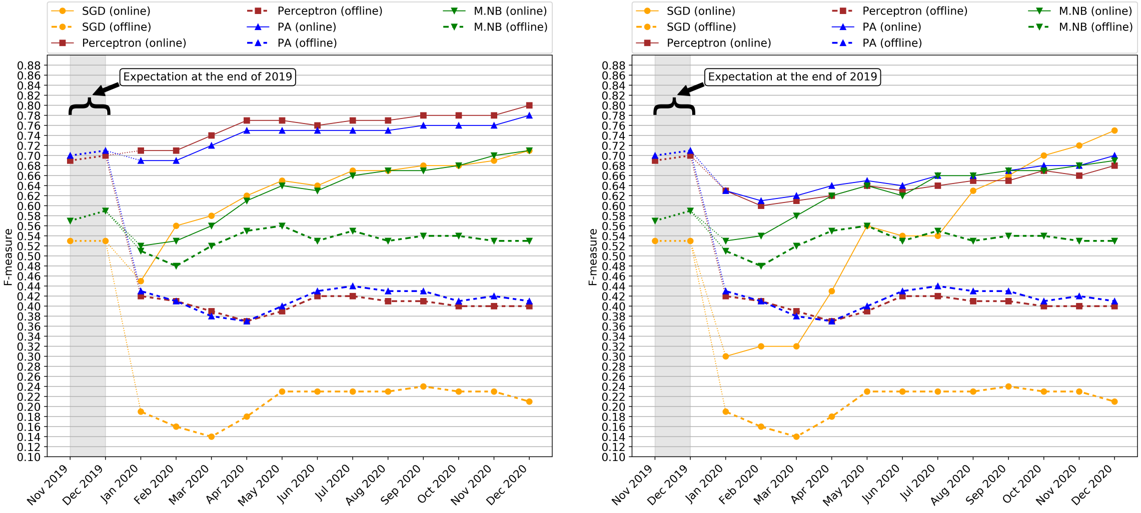


(a) Training only on error.

(b) Training with all the news.

Fig. 6. F-measure obtained in the online learning scenarios with uncertain feedback. The classifier has a random chance of receiving feedback after each prediction, updating the model when feedback is available. Gray area indicates the expected performance considering the last two months of 2019.

feedback in all months but higher than uncertain feedback in January 2020. However, the differences in performance between the types of feedback, in general, are slight. The more significant difference considering the same month was obtained in the experiments with SGD, in March 2020, when the F-measure in the uncertain feedback was 22% lower than that obtained in the immediate feedback. In all other experiments, the difference were smaller. For example, in the experiments with PA, the highest difference in the same month was observed in January 2020, when the F-measure in the uncertain feedback was 5.5% lower than that obtained in the immediate feedback. These results indicate that updating the classifier on error is enough to overcome the concept drift phenomenon. Even in scenarios



(a) Training only on error.

(b) Training with all the news.

Fig. 7. F-measure obtained in the online learning scenarios with delayed feedback. The classifier receives feedback with a delay randomly set between 0 and 20 messages, updating the model when feedback is available.

where the model is updated sporadically, it can adapt to changes in data patterns over time.

5. CONCLUSIONS

Information has a significant impact on people's lives. In historical events, the ease of communication and the speed at which information reaches people can be a powerful resource to help people make good decisions. However, when information is distorted, what was supposed to be an advantage becomes a big problem. For example, the bombing of fake news leaves people confused and frightened. It can greatly negatively influence the results of historical events such as the Covid-19 pandemic and the elections.

The automatic detection of fake news has been a problem that has received several research contributions in recent years. However, most studies evaluate this problem using offline learning methods which have some flaws, such as: (i) they require that all examples should be stored in memory, and; (ii) they suffer from the concept drift phenomenon.

In this study, we evaluated the hypothesis that studies based on the offline paradigm might have overestimated results because we believe the language and historical events are too dynamic to be fit by static learning models. We compared the results obtained by classification methods using offline and online learning paradigms to analyze this hypothesis. We performed experiments with the following state-of-the-art incremental learning method: PA, M.NB, SGD, and Perceptron. They were applied in the classification of news from 2019 and 2020. In the online learning paradigm, we evaluated three different types of feedback: uncertain, delayed, and immediate. We also compared the performance obtained in a scenario where the feedback is presented only when the model makes a prediction error with the ideal scenario when the feedback is given even when the prediction is correct.

Our experiments confirmed the conclusions we obtained in our previous study [Silva and Almeida 2021]. For example, they showed that the performance obtained using the offline learning paradigm degrades over time. Therefore, the studies that use offline learning approaches can present overestimated results. Consequently, the expected performance for the classifiers based on the results obtained in the last months of 2019 was much greater than the reality presented with the 2020 news, where there were two impacting events (Covid-19 and the US presidential elections). These results indicated that the underlying data distribution changed over time and, therefore, we do not recommend using offline models.

Our study also showed that the online learning strategies helped preserve the classification performance over time, even with significant events such as the elections and the Covid-19 pandemic. The incremental update of the predictive model helped the classifiers circumvent the concept drift phenomenon. In conclusion, these results indicated that the online learning paradigm was more appropriate for the fake news data analyzed in this study and also that it is important to consider the chronological order of the news.

Another important point we can observe is that the feedback given only on an error is enough to deal with the concept drift phenomenon. The results obtained in this scenario were competitive with those obtained where the feedback is presented even when the prediction is correct. This result is positive because users generally provide feedback only when the prediction is wrong in real-world applications.

Finally, our study also indicated that the type of feedback is not so crucial for the classifier to adjust to changes in the underlying data distribution over time. The differences between the performance obtained by the classifiers in the immediate, uncertain, and delayed feedback were very small, which is evidence that even in scenarios where the model is updated sporadically, the classifier can overcome the concept drift phenomenon.

In future research, we intend to investigate online learning methods in fake news written in other languages, such as Portuguese and Spanish. We also intend to investigate if the conclusions presented in this study are preserved when combining the bag-of-word representation with linguistic-based features. In another study [Silva et al. 2020], in which we evaluated only the offline learning paradigm, we found potential benefits in combining these two features.

The immediate feedback assumes that the true class of a given document is available immediately after it is classified, which is unrealistic. On the other hand, the uncertain and delayed feedback scenarios are more realistic but can be affected when the user does not provide the true class. In future work, we intend to investigate a "query by committee" feedback strategy applied for fake news detection. In this strategy, an ensemble of classifiers can decide regarding label query [Krawczyk and Woźniak 2017].

In this study, we evaluated three types of feedback, but other variations can also occur in real-world applications. For example, some news cannot be presented to the filter. Moreover, we can consider a delay between the publication of the news and its presentation to the filter. Therefore, future research can assess the impact of these other types of realistic scenarios.

REFERENCES

Alberto, T. C., Lochter, J. V., and Almeida, T. A. Post or block? advances in automatically filtering undesired comments. *Journal of Intelligent & Robotic Systems* 80 (1): 245–259, 2015a.

Alberto, T. C., Lochter, J. V., and Almeida, T. A. Tubespam: Comment spam filtering on youtube. In *Proceedings of the 14th International Conference on Machine Learning and Applications (ICMLA'15)*. IEEE, Miami, FL, USA, pp. 138–143, 2015b.

Allcott, H. and Gentzkowf, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31 (2): 211–236, 2016.

Almeida, T. A., Silva, T. P., Santos, I., and Hidalgo, J. M. G. Text normalization and semantic indexing to enhance instant messaging and SMS spam filtering. *Knowledge-Based Systems* vol. 108, pp. 25–32, May, 2016.

Almeida, T. A. and Yamakami, A. Facing the spammers: A very effective approach to avoid junk e-mails. *Expert Systems with Applications* 39 (7): 6557–6561, June, 2012.

Almeida, T. A., Yamakami, A., and Almeida, J. Spam filtering: how the dimensionality reduction affects the accuracy of naive Bayes classifiers. *Journal of Internet Services and Applications* 1 (3): 183–200, Feb., 2011.

Alves, J. L., Weitzel, L., Quaresma, P., Cardoso, C. E., and Cunha, L. Brazilian presidential elections in the era of misinformation: A machine learning approach to analyse fake news. In *Progress in Pattern Recognition,* *Image Analysis, Computer Vision, and Applications*, I. Nyström, Y. Hernández Heredia, and V. Milián Núñez (Eds.). Springer International Publishing, Cham, pp. 72–84, 2019.

Biesialska, M., Biesialska, K., and Costa-jussà, M. R. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 6523–6541, 2020.

Bittencourt, M. M., Silva, R. M., and Almeida, T. A. ML-MDLText: A multilabel text categorization technique with incremental learning. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS'19)*. IEEE, Salvador, BA, Brasil, pp. 1–6, 2019.

Bittencourt, M. M., Silva, R. M., and Almeida, T. A. ML-MDLText: An efficient and lightweight multilabel text classifier with incremental learning. *Applied Soft Computing* vol. 96, pp. 106699, Nov., 2020.

Cardoso, E. F., Silva, R. M., and Almeida, T. A. Towards automatic filtering of fake reviews. *Neurocomputing* vol. 309, pp. 106–116, 2018.

Charles F. Bond, J. and DePaulo, B. M. Accuracy of deception judgments. *Personality and Social Psychology Review* 10 (3): 214–234, 2006.

Cormack, G. V. Trec 2007 spam track overview. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC'2007)*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, pp. 1–9, 2007.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. Online passive-aggressive algorithms. *Journal of Machine Learning Research* vol. 7, pp. 551–585, Dec., 2006.

Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A., and de Carvalho, A. C. P. L. F. *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, 2021.

Faustini, P. and Ferreira Covões, T. Fake news detection using one-class classification. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems (BRACIS'19)*. IEEE, Salvador, BA, Brazil, pp. 592–597, 2019.

Freund, Y. and Schapire, R. E. Large margin classification using the perceptron algorithm. *Machine Learning* 37 (3): 277–296, Dec., 1999.

Galhardi, C. P., Freire, N. P., Minayo, M. C. d. S., and Fagundes, M. C. M. Fato ou fake? uma análise da desinformação frente à pandemia da COVID-19 no Brasil. *Ciência & Saúde Coletiva* vol. 25, pp. 4201 – 4210, 10, 2020.

Gama, J., Sebastiao, R., and Rodrigues, P. P. On evaluating stream learning algorithms. *Machine Learning* 90 (3): 317–346, Mar., 2013.

Ghosh, S. and Shah, C. Towards automatic fake news classification. *Proceedings of the Association for Information Science and Technology* 55 (1): 805–807, 2018.

Gôlo, M., Caravanti, M., Rossi, R., Rezende, S., Nogueira, B., and Marcacini, R. Learning textual representations from multiple modalities to detect fake news through one-class learning. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. WebMedia '21. Association for Computing Machinery, New York, NY, USA, pp. 197–204, 2021.

Gruppi, M., Horne, B. D., and Adali, S. NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *CoRR* vol. abs/2003.08444, pp. 1–5, 2020.

Gruppi, M., Horne, B. D., and Adali, S. NELA-GT-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *CoRR* vol. abs/2102.04567, pp. 1–6, 2021.

Horne, B. D., Nørregaard, J., and Adali, S. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology* 11 (1): 7:1–7:23, Dec., 2019.

Kaliyar, R. K., Goswami, A., and Narang, P. Multiclass fake news detection using ensemble machine learning. In *2019 IEEE 9th International Conference on Advanced Computing (IACC)*. IEEE, pp. 103–107, 2019.

Kaliyar, R. K., Goswami, A., Narang, P., and Sinha, S. FNDNet – a deep convolutional neural network for fake news detection. *Cognitive Systems Research* vol. 61, pp. 32–44, 2020.

Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., and Iqbal, A. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications* vol. 4, pp. 100032, 2021.

Kolter, J. Z. and Maloof, M. A. Dynamic weighted majority: An ensemble method for drifting concepts. *The Journal of Machine Learning Research* vol. 8, pp. 2755–2790, 2007.

Krawczyk, B. and Woźniak, M. Online query by committee for active learning from drifting data streams. In *2017 International Joint Conference on Neural Networks (IJCNN)*. pp. 2120–2127, 2017.

Ksieniewicz, P., Zyblewski, P., Choraś, M., Kozik, R., Giełczyk, A., and Woźniak, M. Fake news detection from data streams. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8, 2020.

McCallum, A. and Nigam, K. A comparison of event models for naive Bayes text classification. In *Proceedings of the 15th AAAI Workshop on Learning for Text Categorization (AAAI'98)*. AAAI Press/The MIT Press, Madison, Wisconsin, pp. 41–48, 1998.

Minku, L. L., White, A. P., and Yao, X. The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Transactions on Knowledge and Data Engineering* vol. 22, pp. 730–742, 2010.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *13th International Conference on Computational Processing of the Portuguese Language (PROPOR'2018)*. Springer International Publishing, Canela, Rio Grande do Sul, Brazil, pp. 324–334, 2018.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3391–3401, 2018.

Pérez-Rosas, V. and Mihalcea, R. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 440–445, 2014.

Pérez-Rosas, V. and Mihalcea, R. Experiments in open domain deception detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1120–1125, 2015.

Rasool, T., Butt, W. H., Shaukat, A., and Akram, M. U. Multi-label fake news detection using multi-layered supervised learning. In *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*. ICCAE 2019. Association for Computing Machinery, New York, NY, USA, pp. 73–77, 2019.

Rezayi, S., Balakrishnan, V., Arabnia, S., and Arabnia, H. R. Fake news and cyberbullying in the modern era. In *Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence*. CSCI 2018. IEEE, New York, NY, USA, pp. 7–12, 2018.

Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24 (5): 513–523, Aug., 1988.

Salvi, C., Iannello, P., Cancer, A., McClay, M., Rago, S., Dunsmoor, J. E., and Antonietti, A. Going viral: How fear, socio-cognitive polarization and problem-solving influence fake news detection and proliferation during covid-19 pandemic. *Frontiers in Communication* vol. 5, pp. 127, 2021.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19 (1): 22–36, 2017.

Silva, R. M. and Almeida, T. A. How concept drift can impair the classification of fake news. In *Proceedings of the 9th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe'21)*. Brazilian Computing Society, Rio de Janeiro, RJ, Brazil, pp. 1–8, 2021.

Silva, R. M., Almeida, T. A., and Yamakami, A. MDLText: An efficient and lightweight text classifier. *Knowledge-Based Systems* vol. 118, pp. 152–164, Feb., 2017.

Silva, R. M., de Sales Santos, R. L., Pardo, T. A. S., and Almeida, T. A. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications* vol. 146, pp. 1–48, May, 2020.

Song, C., Ning, N., Zhang, Y., and Wu, B. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management* 58 (1): 102437, 2021.

van der Linden, S., Roozenbeek, J., and Compton, J. Inoculating against fake news about covid-19. *Frontiers in Psychology* vol. 11, pp. 1–7, 2020.

Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science* 359 (6380): 1146–1151, 2018.

Wang, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pp. 422–426, 2017.

Zarocostas, J. How to fight an infodemic. *The lancet* 395 (10225): 676, 2020.

Zhang, S. and Kejriwal, M. Concept drift in bias and sensationalism detection: An experimental study. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM'19. Association for Computing Machinery, New York, NY, USA, pp. 601–604, 2019.

Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21th International Conference on Machine Learning (ICML'04)*. ACM, Banff, Alberta, Canada, pp. 116–123, 2004.

Zhou, L., Burgoon, J., Twitchell, D., Qin, T., and Nunamaker Jr., J. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems* 20 (4): 139–165, 2004.

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice* 1 (2): 12:1–12:25, June, 2020.

Zhou, X. and Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys* 53 (5): 109:1–109:40, 2020.