SAMS: Human-in-the-loop Approach to Combat the Sharing of Digital Misinformation

Shaban Shabani[a,b], Zarina Charlesworth[b], Maria Sokhn[b] and Heiko Schuldt[a]

[a]Department of Mathematics and Computer Science, University of Basel, Switzerland
[b]Institut de Digitalisation, University of Applied Sciences Western Switzerland (HES-SO), Neuchâtel, Switzerland

Abstract

Spread of online misinformation is an ubiquitous problem especially in the context of social media. In addition to the impact on global health caused by the current COVID-19 pandemic, the spread of related misinformation poses an additional health threat. Detecting and controlling the spread of misinformation using algorithmic methods is a challenging task. Relying on human fact-checking experts is the most reliable approach, however, it does not scale with the volume and speed with which digital misinformation is being produced and disseminated. In this paper, we present the SAMS Human-in-the-loop (SAMS-HITL) approach to combat the detection and the spread of digital misinformation. SAMS-HITL leverages the fact-checking skills of humans by providing feedback on news stories about the source, author, message, and spelling. The SAMS features are jointly integrated into a machine learning pipeline for detecting misinformation. First results indicate that SAMS features have a marked impact on the classification as it improves accuracy by up to 7.1%. The SAMS-HITL approach goes one step further than the traditional human-in-the-loop models in that it helps raising awareness about digital misinformation by allowing users to become self fact-checkers.

Keywords

Digital Misinformation, Machine Learning, Crowdsourcing, Human-in-the-Loop

1. Introduction

Advances in mobile technology have allowed for an unprecedented spread of information and both mis- and disinformation. The ease of transmission and sharing, the use of social media and messaging apps coupled with the increasing penetration of the Internet, provides a fertile ground for its spread [1]. As pointed out by Ciampaglia [2], the risk is the massive, uncontrolled, and often systemic spread of untrustworthy content.

Digital misinformation comes in a variety of forms from entirely false to the integration of one or two misleading sentences in a piece of real news or just a provocative misleading title in introduction to a correct piece of news. In addition to this, one also finds rumor, hoaxes, satire, and conspiracy theories contributing to what can be characterized as an online false information ecosystem [3]. One can group such information under the umbrella term of misinformation. More recently, one also sees a distinction between misinformation, which can be spread with or without intent to mislead, and disinformation, which intends to spread false information [1].

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) – Stanford University, Palo Alto, California, USA, March 22-24, 2021.

EMAIL: shaban.shabani@unibas.ch (S. Shabani); zarina.charlesworth@he-arc.ch (Z. Charlesworth); maria.sokhn@hes-so.ch (M. Sokhn); heiko.schuldt@unibas.ch (H. Schuldt)
ORCID: 0000-0003-4710-6091 (S. Shabani); 0000-0002-2898-5716 (Z. Charlesworth); 0000-0001-7586-0564 (M. Sokhn); 0000-0001-9865-6371 (H. Schuldt)

The problem with misinformation is that it is pervasive and runs through all types of media from print to radio to online. The latter grew considerably during 2016 American presidential election and with the onset of the COVID-19 pandemic in March of 2020 has now taken on alarming proportions. In the words of T. A. Ghebreyesus[1], director-general of the WHO, speaking of COVID-19: *"We are not just fighting an epidemic; we are fighting an infodemic"*. Digital misinformation spreads faster and more easily than this virus, and is just as dangerous. Unlike the virus, however, COVID-19 related news has two strains *true* and *false* with the latter inundating social media channels, and going largely unverified. The expression *infodemic* was first used in 2003 by Rothkopf [4] when writing about the SARS epidemic and highlighting the negative impact that misinformation had on controlling the then health crisis – a crisis far from the size of what we are now experiencing with COVID-19. In today's COVID-19 influenced world, we are indeed dealing with an infodemic and the question is how best to control it and fight the spread of misinformation. In order to counter this, and in light of the high level of user distrust towards online fact-checking services, there is an urgent need to train individuals to evaluate the veracity of the information they are receiving and sharing, and give them the tools to become fact-checkers in their own right.

In recent years, fact-checking services have become the norm for journalists and are now also easily accessible by the public. Although the majority address issues in the political arena, SNOPES[2], a well known service, started out primarily by debunking urban legends. Such services certainly have a role to play in response to the challenge of online misinformation [5, 6], however, there is increasing interest in coming up with an automated and scalable response [7]. Despite the availability of fact-checking services, research suggests that there is a high level of distrust for such services [8, 9]. Yet another argument in support of the development of an individual user application.

This research focuses specifically on digital misinformation. As this is technology-based, the solutions considered are also often only technology-related [7]. Some advocate that the response to the online spread of misinformation is through technology [10] and the integration of Artificial Intelligence (AI), others lean more towards human fact checkers. An alternate possibility is through a combination of the two allowing for the development of a high level performing model used by individuals. Research in the area of mixed-initiative fact-checking [11, 12, 13] suggests that AI alone cannot be as accurate as when the human element is integrated in the fact-checking process. Human-in-the-loop AI (HAI) systems face different challenges in terms of effective performance due to fact that individuals are involved. It is, however, possible to train models to a significant level of accuracy.

We suggest going one step further in the battle to slow-the-flow by taking an HAI approach as well as involving those who are at the source by getting them on board as fact-checkers in their own rights. In order to do this, we have developed a user friendly tool that both identifies the veracity of the news and calls on the user to self-check four critical indicators on their own. Our proposed framework is called SAMS. Following a review of the published research and literature on credibility indicators [14, 15, 16] and fact-checking guides [9, 17] the choice of a limited number of checks seemed to be appropriate.

In order to be best armed to counter the spread of misinformation, it is important to see who is spreading it. Spring [18] suggests grouping the spreaders into seven categories ranging from the "Joker" to the "Politician" and including "well-intentioned family members". Zannettou et al. [3] go one step further including even bots for a total of ten categories. Keeping in mind the fact that we were looking for a limited number of indicators and yet ones that could be applied to all categories, the ones that repeatedly came to the top of the list were: *source*, *author*, *message*, and *spelling* (SAMS).

[1] https://www.who.int/dg/speeches/detail/munich-security-conference

[2] https://www.snopes.com/fact-check/

We went with these and launched into the development of a prototype.

In this paper, we propose and evaluate the SAMS-HITL method. It uses supervised machine learning models to classify news articles into *false* and *true*. It leverages the fact-checking skills of humans by providing answers to SAMS related questions of the news articles. The human feedback about the four SAMS indicators is joined with the automatically extracted features from the text of news articles. We evaluate this approach on a recently published dataset with articles related to COVID-19. Preliminary results show that the SAMS human-in-the-loop (SAMS-HITL) approach outperforms methods that rely only on automated techniques. Feeding the model with information about the *source*, *author*, *message*, and *spelling* provides higher accuracy in the classification task.

Our contributions can be summarized as follows:

- We conceptualized SAMS-HITL as a user friendly option to check information, which calls on human input and is backed up by machine learning models
- We designed a crowdsourcing task that leverages the human cognitive skills of online crowd workers on providing answers to SAMS questions, using an aggregation method to infer the true answer based on multiple judgments
- We implemented and evaluated SAMS-HITL approach on a dataset with COVID-19 related news articles. The proposed technique performs better than automatic classification models. Preliminary results indicate that SAMS features have a marked impact on the classification as it improves accuracy by up to 7.1%.

The remainder of the paper is structured as follows: Section 2 introduces the concept behind SAMS and describes the individual components. Section 3 introduces the dataset and describes the implementation. Section 4 details the evaluation and results of our methods. Section 5 presents related work and Section 6 concludes with suggestions for further research.

2. Concept

In this section, we present the components of the SAMS-HITL approach. Figure 1 illustrates the overall architecture of SAMS.

2.1. Machine Learning Component

Considering the rapid growth of online data and spread of misinformation, and the high impact it has on the society, efficient and effective data processing tools are essential. Approaches based on machine learning and deep learning techniques [19] have been comprehensively considered for fake news detection. The core component of SAMS is the supervised machine learning model which analyzes the news content. This model consists of two phases: i) *feature extraction*, and ii) *model construction*.

Feature extraction is performed on the text coming from the headline and the body text. The headline of a news item is a short text that is meant to catch the attention of the reader, whereas the body text is the main part that details the news story. We consider two types of features: *statistical* and *sentiment* features based on linguistic characteristics. The statistical features are extracted using the Term-Frequency Inverse-Document-Frequency (tf-idf) algorithm which measures the importance of words in the text document. Analyzing the sentiment of the news is very important especially when taking into account that much of the misinformation being spread started out as disinformation with the intent to deceive rather than to report objectively, sometimes for political or financial gain, and

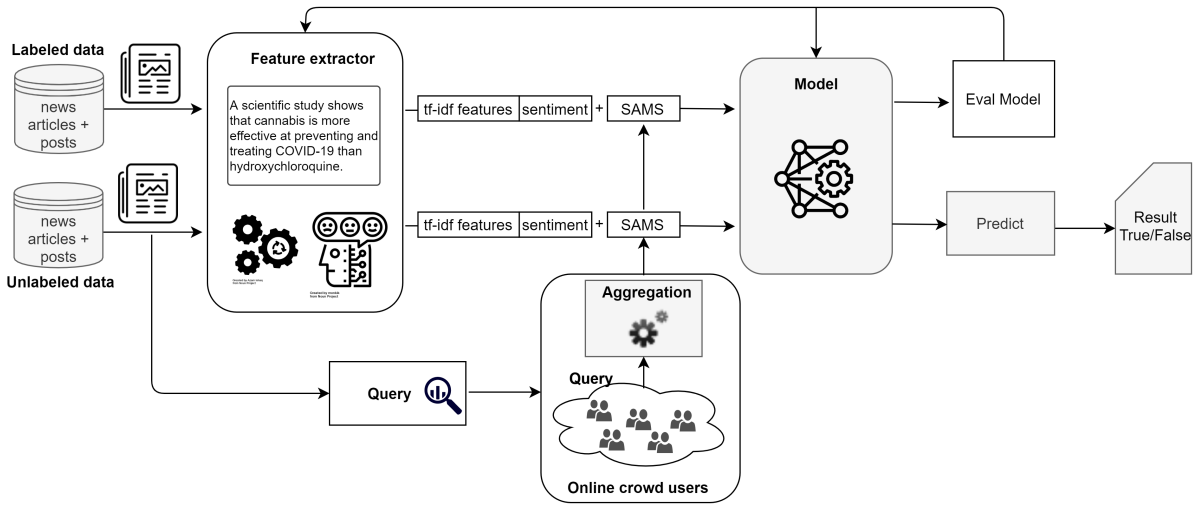

Figure 1: SAMS overall architecture – the feature extractor module initially performs a text cleaning step and generates the tf-idf features together with sentiment features. The labeled data includes new articles that are categorized as false or true. Retrieving SAMS features is done by querying the crowdsourcing module and collecting answers from multiple users. Judgments from users are aggregated and injected to the set of automatically extracted features (tf-idf and sentiment) and finally the combined feature set is used by the model for prediction.

in the COVID-19 pandemic situation to exploit public fear and uncertainty. Therefore, the sentiment features focus on capturing the *objectivity* aspect, the *mood*, *modality*, and *polarity* of the reported news. Additionally, the length of news stories is an important aspect as misinformation generated on social network channels tends to be short and catchy.

Model construction builds the machine learning model to perform the classification, in order to differentiate between false and true news. For the evaluations, we selected four different state-of-the-art algorithms: Logistic Regression (LOG), Random Forest (RF), Gradient Boosting Classifier (GBC), and Support Vector Machines (SVM).

2.2. SAMS – Source, Author, Message, Spelling

Classifying news articles by relying solely on machine learning models based on news content is a challenging task. One reason is that spreaders of misinformation have advanced their writing style and the language used in the news with the aim of distorting the truth and bypass detection by style-based models. Another very important factor is the length of the news: false stories, especially in the era of the COVID-19 pandemic, tend to be short and alerting, making it difficult to automatically analyze the message conveyed by such stories. Research in the area of mixed-initiative fact-checking [11, 12, 13] suggests that machine learning alone cannot be as accurate as when the human element is integrated in the fact-checking process. We have identified that important features when performing fact-checking are: *source*, *author*, *message*, and *spelling*. Answering the questions about the four features of SAMS automatically is non-viable. In contrast, humans have the potential to do better, as they can perform fact-checking skills, by searching for facts on trustworthy data resources. We define a process with tips and tricks to easily answer each of the questions for SAMS.

Source – taking a critical look at the source, both data and metadata, is the first step. The goal is to understand if the news stories have sources and if the sources are reliable. To better evaluate if the

sources are trustworthy, we take a look at where the information originated, inspect if the references are stated and if so, trace the references checking if they are correct and trustworthy.

Author – in principle, real and serious news articles always have an author. Therefore, the first step is to identify if there is an author of the news item. If so, further inspections include if the author is a journalist, their affiliation, academic or professional credentials. Furthermore, a check for related publications by the same author can be made.

Message – the message should be clear, balanced, and unbiased. Guidelines to identify misinformation suggest checking for unsupported or outrageous claims, if there is a push to share the information, lack of quotes, references or contributing sources, and identifying if the headlines provoke strong emotions.

Spelling – reputable sources will proofread material prior to publishing. Misinformation tends to have grammar mistakes such as repeated spelling mistakes, poor grammar, incorrect punctuation, use of different fonts, the writing of entire words or phrases written in capital letters, etc.

2.3. Human-in-the-Loop Approach

Obtaining reliable information related to the four aspects of SAMS is crucial for our approach towards misinformation detection. While experts such as journalists are well trained to search for the right data sources to find facts, employing them becomes expensive. Considering the amount and velocity of potential disseminated digital misinformation, this approach is not scalable in terms of time.

On the other hand, crowdsourcing has been widely deployed for small tasks as it leverages the collective human skills of extensive online crowd worker communities. In specific scenarios, crowdsourcing has shown to be an alternative service to replace experts with specific domain knowledge for labelling. In this work, we design a crowdsourcing component which aggregates the inputs from multiple users to infer the true answers related to SAMS questions. The output of the crowd answers is a vector of four binary values, each value corresponding to the SAMS questions. As could be seen in Figure 1, the output is encoded into a binary feature vector which later is appended to the feature vector generated using the tf-idf algorithm and the sentiment features described in Section 2.1. Finally, the concatenated feature vector is used for training and evaluating the machine learning models.

3. Dataset and Implementation

In this section, after presenting the dataset we used to evaluate our approach, we describe the implementation of the SAMS pipeline.

3.1. Dataset

In this experimental setup, we use the CoAID dataset collected and annotated by Cui and Lee [20]. The dataset consists of *true* and *false* news about COVID-19 from diverse sources mainly covering websites and social network platforms. There are several types of entries such as "news articles" collected from fact-checking reliable sources, "claims" posted by official channels of WHO, "user engagement" which include twitter posts and replies, and other "social platform posts" such as Facebook, Instagram, etc. For each entry, there is the title, abstract, content, keywords, and URL of the article or the post. Our interest is in analyzing news posts that contain potentially longer text and posted on various social media channels (online newspapers, blogs, communication apps etc.), therefore we focus only news articles, skipping the twitter posts with short text. As a result, we filtered the false and true news from the CoAID dataset, ending up with 1,127 *true* samples, and 266 *false* samples. Considering that

the two classes are imbalanced, finally from the dataset we selected all 266 available false entries and 264 randomly sampled real news articles. Figure 2 illustrates the distribution of news length by word count and Table 1 describes the descriptive statistics of the articles length. Average length of true and false news articles is 35 and 33 words, respectively.

Class	Min	Max	Mean	StDev	Med	Total
false	7	85	33.27	16.05	31	8'850
true	11	89	35.17	12.32	34	9'285

Table 1: Descriptive statistics of the dataset articles word length

Features	LOG	SVM	GBC	RF
TF	76.4	77.4	75.2	77.9
TFS	82.2	80.6	87.6	91.5
TFST	87.8	82.8	93.1	93
TFSC	87.1	83.7	94.7	93.6

Table 2: Classification results (f1 score)

3.2. Implementation

The first step towards implementing a model is data pre-processing. Since this is a text classification task, text cleaning is a useful process and that includes removing frequent words that provide non-unique information to the model (stop words) and special characters, and applying stemming and lemmatization. This part is important for extracting features using the tf-idf technique. On the other hand, extracting sentiment features is possible on the raw text and this is done with the pattern.en tool [21]. The sentiment features include: i) *polarity* which is given as a value between -1.0 (completely negative) and +1.0 (completely positive); ii) *subjectivity* which is a value between 0.0 (very objective) and 1.0 (very subjective); iii) *modality* feature represents the degree of certainty as a value between -1.0 and +1.0, where values higher than +0.5 represent facts ; iv) *mood* feature is a categorical value based on auxiliary verbs and the answer can be either "indicative", "imperative", "conditional", or "subjunctive". Additionally, we add the word count to the feature vector.

The aim of this work is evaluating the SAMS-HITL approach and the importance of the four indicators in the classification task. Doing that, called for having answers to the four SAMS questions for every news record from the dataset. Initially, we labelled manually the 530 dataset entries. A trained annotator used an in-house developed web annotation interface to answer the SAMS questions for each dataset entry. The annotation was a tedious task that took approximately 30 hours. After that, we designed a crowdsourcing job on the Microworkers[3] crowdsourcing platform. Each news story was used to generate a Human Intelligence Task (HIT), asking online crowd participants to provide the answers to SAMS questions which were stated as follows:

1. *Is there a source in this news article?* – Yes/No.
2. *Is there an author in this news article?* – Yes/No.
3. *Is the message of this news article clear, unbiased, and balanced?* – Yes/No.
4. *Is spelling correct on this news article?* – Yes/No.

A HIT contained the URL of the original news story, the headline, and the body text. Crowd workers were instructed to click on the news link, inspect and analyze the article always considering the four questions that they were asked to answer. We used data quality control mechanisms [22] such as redundancy where for each task we asked three workers from three different regions: USA, Europe, and Asia. Analysis on demographics and dynamics of crowd workers on crowdsourcing platforms has shown that these regions are mostly represented [23] and this is the case in the Microworkers platform as well. Additionally, collecting judgments from crowd workers from different regions could be an

[3] https://www.microworkers.com

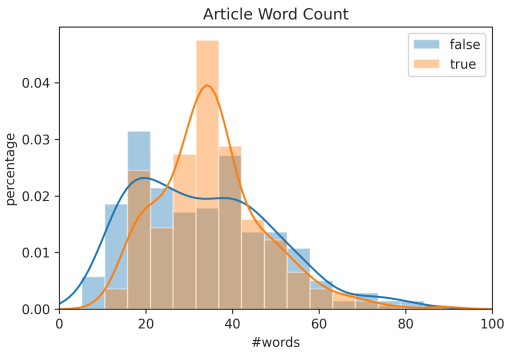

Figure 2: Distribution of news length by word count

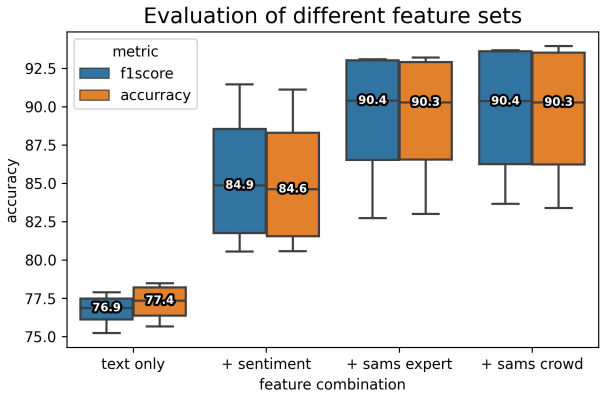

Figure 3: Performance impact of different set of features

important factor as diversity matters when it comes to label quality [24]. Furthermore, task design techniques are important element for high quality data, therefore guidelines with tips and examples were part of the instructions in the crowdsourced job.

The choice of obtaining multiple judgments for the same question from different user demographics can increase the quality of crowdsourced data. Depending on the task complexity, various aggregation techniques can be applied, such as voting strategies, profile-based or iterative aggregation algorithms [25]. The majority voting is the simplest as it is non-iterative and does not require pre-processing, aggregating each object independently by choosing the label with highest votes. In our scenario, for a single HIT that has SAMS questions, we would get three responses for each of the four questions and the aggregation selects the answers with highest votes. Since crowd workers can have different levels of expertise, profile-based strategies take into account information from their past contributions to build a ranking score, as well incorporate additional information such as location, domain of interests etc. The reputation score of crowd workers can be updated dynamically based on their performance on the existing task. Iterative algorithms are based on a sequence of computational rounds where in each round the probabilities of possible labels for each object are computed and updated repeatedly until convergence. For the answer aggregation in our SAMS-HITL approach, we applied the Dawid and Skene algorithm [26], a model that is based on the Expectation-Maximization (EM) principle to model the worker's reliability with a confusion matrix.

4. Evaluation

In this section, we outline the evaluation of the proposed SAMS-HITL approach. We used a 10-fold cross validation for evaluating the performance of the models, and accuracy and f1 score as evaluation metrics. To analyze the evaluation of models and the importance of the three different sets of features described in Sections 2.1 and 2.2, we run the evaluation of the models for each setting separately. We consider the following combinations of features:

(i) tf-idf features (TF)
(ii) tf-idf + sentiment features (TFS)
(iii) tf-idf + sentiment features + SAMS trained annotator features (TFST)
(iv) tf-idf + sentiment features + SAMS crowd features (TFSC)

Table 2 shows the f1 score of the models under different features setup. When considering only tf-idf (TF) features extracted from the headline and body text, the Random Forest model achieves the highest accuracy of 77.9%. Appending the sentiment features (TFS) showed to have considerable impact on the performance of the models, lifting up the accuracy to 91.5%. Finally, adding the SAMS features in both options with crowdsourcing (TFSC) and trained annotator (TFST) shows a positive impact on the models' performances. The Gradient Boosting classifier model achieves the highest accuracy of 94.7% with SAMS features obtained by aggregating the answers from the crowd. Interestingly, the TFSC approach performs slightly better than TFST in three out of the four models. It can be observed that sentiment features (TFS) have distinct impact on the accuracy for the Random Forest model compared to TF features, improving the accuracy by 13.6%. The TFSC features increase the accuracy by another 2.1% . This directs us to further inspect the evaluation with more samples and test the models performance with additional data sources. In such a scenario, we would expect that both SAMS (TFST and TFSC) features will make an even larger difference compared to the other settings.

Figure 3 shows the effect of the four combinations of features during the evaluation. We can see that sentiment features (TFS) have significant impact on the performance compared to the tf-idf (TF) features. Furthermore, we can observe that both combinations with SAMS features overall indicate a significant difference compared to the TFS and TF approaches. Additionally, we evaluate the importance of features with a tree of forests and analysis shows that most significant features are the SAMS features, where feature about the *message* had highest score, followed by spelling, source, and author. From the sentiment features, modality and the text length (word count) appeared in the top ten list. Finally, Pearson's correlation analysis shows that *source* feature has a moderate positive correlation of 0.57 with the class, followed by *message* at 0.51.

5. Related Work

In recent years, automatic misinformation classification has been extensively used under specific supervised scenarios. Shu et al. [19] explore the characterization of fake news in social media and the data mining perspectives. As an emerging topic, misinformaton has drawn attention of the research communities in different disciplines. As a result, several datasets have been published, related to political news [27], rumor debunking [28], fake vs. satire detection [29], FEVER dataset for verification of textual sources [30], and a more recent dataset with news related to the COVID-19 pandemic [20]. Significant efforts have been made on exploring the potential of deep learning [27, 31] and the linguistic and semantic aspects [32, 33].

Crowdsourcing as a methodology can assist in classification of news articles by fact-checking the statements. Tschiatschek et al. [34] propose a strategy of flagging news considered as false by social network users, and deploys an aggregation method to select subset of those flagged news for evaluation by experts. A recent work by Roitero et al. [35] analyzed how the crowd users assessed the truthfulness of false and true statements related to COVID-19 pandemic. Results show that the crowd is able to accurately classify statements and achieve a certain level of agreement with expert judgments. In response to combating the COVID-19 misinformation, Li et al. [36] developed the Jennifer chatbot which helps users to easily find information related to COVID-19. The chatbot provides reliable sources and diverse topics maintained by a global group of volunteers.

Considering the sensitivity and the risk of misinformation spreading on one hand, and the limitations of both automated and human-based methods, hybrid human-machine approaches have been envisioned [12, 37]. For instance, the hybrid machine-crowd [38] approach has demonstrated higher

accuracy for classification of fake and satiric stories. It uses a high-confidence switching method where crowd feedback is requested whenever the ensemble of machine learning models fails to achieve unanimity and high accuracy. A hybrid human-machine interactive approach [11] based on a probabilistic graphical model combines machine learning model predictions with crowd annotations for fact-checking. The follow-up user study [39] shows that predictions from automated models can help users in assessing claims correctly, hence tending to trust the system, even when model predictions were wrong. However, enabling interaction and having transparent model predictions has the potential of training the users to build their own skills for fact-checking.

6. Conclusion and Future Work

In this paper, we have addressed the issue of classifying news stories related to the COVID-19 pandemic. We presented SAMS-HITL framework for misinformation detection. SAMS-HITL combines statistical and sentiment based features automatically extracted from the text of the news articles with the features related to the source, message, author, and spelling of the article obtained via crowdsourcing. Preliminary results showed that the four SAMS features are the most important features of the classification model and it has high impact on the overall classification accuracy. In summary, our proposed framework leverages the efficiency of machine learning models over a large amount of data and the quality of human intelligence for fact-checking. This method is helpful for social networks which could benefit from the high availability of their platform members and leverage their fact-checking skills to provide feedback on SAMS questions on news articles that are posted and shared on their platform. The SAMS-HITL approach goes one step further than the traditional HAI models in that it calls on the users to answer the four questions themselves thus raising user awareness about digital misinformation. The SAMS-HITL prototype application is currently being developed. Our objective is to help users check news articles, time train them to have a critical view and raise awareness about misinformation. In the long run, the impact can only be positive as even without the SAMS-HITL app people will think twice before passing news along.

A limitation in the results presented is the size of the dataset and the potential bias in the classes due to the limited diversity of sources and the length of the text in the news articles. Validating SAMS-HITL will call for its application on a much larger dataset. Having more data gives opportunity to consider word-embedding techniques for feature extraction and application of deep learning models for classification. Future work intends to further investigate the SAMS features. One direction is to explore the potential of automatically answering the questions related to author and spelling, which could reduce the human effort. Automated tools in combination with a customized text processing algorithm can be used to identify grammar mistakes and generate a score for the spelling. For news articles published on web portals, identifying and extracting metadata information related to the author could be done automatically. However, this is challenging for news stories published and shared via different social network channels. Further work on SAMS features will investigate the impact of using a score range instead of the binary yes/no values.

Acknowledgments

This work was partly funded by the HES-SO (project no. 104353) and Hasler Foundation in the context of the project City-Stories (contract no. 17055).

References

[1] L. Ha, L. A. Perez, R. Ray, Mapping recent development in scholarship on fake news and misinformation, 2008-2017: Disciplinary contribution, topics, and impact, American Behavioral Scientist August (2019) 1–26.

[2] G. L. Ciampaglia, Fighting fake news: a role for computational social science in the fight against digital misinformation, Journal of Computational Social Science 1 (2018) 147–153.

[3] S. Zannettou, M. Sirivianos, J. Blackburn, N. Kourtellis, The web of false information: Rumors, fake news, hoaxes, clickbait , and various other shenanigans, Journal of Data and Information Quality 11 (2019) 37.

[4] D. J. Rothkopf, When the buzz bites back, The Washington Post (2003) BO1.

[5] J. R. Harman, Collateral Damage: The Imperiled Status of Truth in American Public Discourse and Why it Matters to You, Author House, Bloomington, Indiana, 2014.

[6] P. B. Brandtzaeg, A. Følstad, M. A. C. Dominguez, How journalists and social media users perceive online fact-checking and verification services, Journalism Practice 12 (2018) 1109–1129.

[7] L. Graves, Understanding the Promise and Limits of Automated Fact-Checking, Report, Reuters Institute for the Study of Journalism et University of Oxford, 2018.

[8] P. B. Brandtzaeg, A. Følstad, Trust and distrust in online fact-checking services, Communications of the ACM 60 (2017) 65–71.

[9] Evaluating resources: the crapp test, 2015. URL: https://researchguides.ben.edu/c.php?g=261612&p=2441794.

[10] L. Konstantinovskiy, O. Price, M. Babakar, A. Zubiaga, Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, 2020.

[11] A. T. Nguyen, A. Kharosekar, S. Krishnan, S. Krishnan, E. Tate, B. C. Wallace, M. Lease, Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking, in: UIST '18: The 31st Annual ACM Symposium on User Interface Software and Technology, 2018, pp. 189–199.

[12] G. Demartini, S. Mizzaro, D. Spina, Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities, The Bulletin of the Technical Committee on Data Engineering 43 (2020) 65–74.

[13] G. Rehm, An infrastructure for empowering internet users to handle fake news and other online media phenomena, in: International Conference of the German Society for Computational Linguistics and Language Technology, Springer, 2017, pp. 216–231.

[14] D. Esteves, A. J. Reddy, P. Chawla, J. Lehmann, Belittling the source: Trustworthiness indicators to obfuscate fake news on the web, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 2018, pp. 50–59.

[15] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al., A structured response to misinformation: Defining and annotating credibility indicators in news articles, in: The Web Conference 2018, 2018, p. 603–612.

[16] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al., The science of fake news, Science (2018).

[17] Library guides: News literacy, Accessed 2020. URL: https://nwtc.libguides.com/news.

[18] M. Spring, Coronavirus: The seven types of people who start and spread viral misinformation, BBC Trending (2020). URL: https://www.bbc.com/news/blogs-trending-52474347.

[19] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, SIGKDD Explor. Newsl. 19 (2017).

[20] L. Cui, D. Lee, CoAID: COVID-19 Healthcare Misinformation Dataset, https://arxiv.org/abs/2006.00885, 2020.

[21] T. De Smedt, W. Daelemans, Pattern for python, The Journal of Machine Learning Research 13 (2012) 2063–2067.
[22] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, S. Dustdar, Quality control in crowdsourcing systems: Issues and directions, Internet Computing (2013).
[23] D. Difallah, E. Filatova, P. Ipeirotis, Demographics and dynamics of mechanical turk workers, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 135–143.
[24] G. Kazai, J. Kamps, N. Milic-Frayling, The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, 2012.
[25] N. Quoc Viet Hung, N. T. Tam, L. N. Tran, K. Aberer, An evaluation of aggregation techniques in crowdsourcing, in: Web Information Systems Engineering – WISE 2013, 2013, pp. 1–15.
[26] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, Journal of the Royal Statistical Society: Series C (Applied Statistics) 28 (1979) 20–28.
[27] W. Y. Wang, "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, 2017.
[28] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2016, pp. 1163–1168.
[29] J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett, et al., Fake news vs satire: A dataset and analysis, in: Proceedings of the 10th ACM Conference on Web Science, 2018, pp. 17–21.
[30] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018, pp. 809–819.
[31] N. Ruchansky, S. Seo, Y. Liu, Csi: A hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 797–806.
[32] V. L. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the second workshop on computational approaches to deception detection, 2016, pp. 7–17.
[33] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018.
[34] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, A. Krause, Fake news detection in social networks via crowd signals, in: The Web Conference 2018, WWW '18, 2018, p. 517–524.
[35] K. Roitero, M. Soprano, B. Portelli, D. Spina, V. Della Mea, G. Serra, S. Mizzaro, G. Demartini, The covid-19 infodemic: Can the crowd judge recent misinformation objectively?, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020.
[36] Y. Li, T. Grandison, P. Silveyra, A. Douraghy, X. Guan, T. Kieselbach, C. Li, H. Zhang, Jennifer for COVID-19: An NLP-powered chatbot built for the people and by the people to combat misinformation, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, 2020.
[37] S. Ahmed, K. Hinkelmann, F. Corradini, Combining machine learning with knowledge engineering to detect fake news in social networks-a survey, in: AAAI'19 Spring Symposium, 2019.
[38] S. Shabani, M. Sokhn, Hybrid machine-crowd approach for fake news detection, in: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), 2018, pp. 299–306.
[39] A. T. Nguyen, A. Kharosekar, M. Lease, B. C. Wallace, An interpretable joint graphical model for fact-checking from crowds., in: AAAI Conference on Artificial Intelligence, 2018, pp. 1511–1518.