TTO 2021

Proceedings of the 2021 Truth and Trust Online Conference (TTO 2021)

October 7-8, 2021
Virtual

Production and Manufacturing by
Hacks Hackers
c/o Fletcher Heald and Hildreth PLC
1300 17th St. North, 11th floor
Arlington, VA 22209 United States

Order copies of this TTO proceedings from:

Hacks Hackers
c/o Fletcher Heald and Hildreth PLC
1300 17th St. North, 11th floor
Arlington, VA 22209 United States
admin@truthandtrustonline.com

Preface

The annual Conference for Truth and Trust Online (TTO) took place on October 7-8, 2021. Due to the COVID-19 pandemic, it was held online.

The mission of TTO, now in its third edition, is to bring together all parties working toward improving the truthfulness and trustworthiness of online communications. TTO is an annual forum for academia, industry, non-profit organizations, and other stakeholders to discuss the problems facing (social) media platforms and technical solutions to understand and address them. It is organised as a unique collaboration between practitioners, technologists, academics and platforms, to share, discuss, and collaborate on useful technical innovations and research in the space.

The aim of TTO is to be a forum from which many parties can benefit: (i) for academics, to learn about the real problems that industry is facing and how their proposed solutions can be more impactful, (ii) for industry, to improve their product safety by brainstorming collective actions together with other stakeholders, and (iii) for the public, to gain insights into how their concerns on social media safety are being addressed.

We invited submissions on topics such as misinformation, disinformation, trustworthiness of COVID-19 news and guidance, hate speech, online harassment and cyberbullying, credibility, hyper-partisanship and bias, image/video verification, fake amplification, fake reviews, polarization and echo chambers, transparency in content and source moderation, and privacy requirements.

We invited two kinds of submissions: technical papers and talk proposals. The idea was to attract both fully worked papers and "ideas." The technical papers were an opportunity for authors to publish and present new research, and they are included in these proceedings. In contrast, the talks do not appear in the proceedings and they were designed as an opportunity for scholars, activists, developers, lawyers, ethics experts, fact-checkers, public servants, journalists, and all around researchers to present and discuss ideas related to the topic of the conference.

We received 35 submissions for technical papers, and we accepted 7 of them, an acceptance rate of 20%. Of those. we invited 3 of them to submit an extended version to the ACM Journal of Data and Information Quality special issue on TTO. We further received 35 talk proposals, and we accepted 14 of them, an acceptance rate of 56%.

Here are some statistics about the authors of the paper and talk submissions by country: 75 from USA, 32 from UK, 12 from France, 9 from Switzerland, 8 from Italy, 6 from Qatar, 6 from Germany, 6 from Japan, 5 from Australia, 5 from Israel, 3 from Greece, 3 from South Korea, 3 from Turkey, 2 from Brazil, 2 from Canada, 2 from Luxembourg, 2 from Spain, 1 from Belgium, 1 from Chile, 1 from China, 1 from Hong Kong, 1 from India, 1 from Sri Lanka, 1 from Sweden, and 1 from Tunisia.

Each paper received 3 reviews, and each talk proposal received 2 reviews. There were 34 PC members, and the acceptance decisions were made after intense online discussions between the reviewers and the PC chairs.

We thank the reviewers for their hard work, and to the authors for contributing their very interesting research and discussions, which have allowed us to produce a very interesting and balanced programme.

Isabelle Augenstein and Paolo Papotti,
TTO-2021 PC Chairs

Organization

This conference is a truly unique collaboration between academia, industry, and practitioners.

Organizers

General Chairs:

Baybars Örsek (International Fact-Checking Network / Poynter Institute)
Christos Christodoulopoulos (Amazon)

Program Chairs:

Isabelle Augenstein (University of Copenhagen, CheckStep)
Paolo Papotti (EURECOM)

Publication Chair:

Dustin Wright (University of Copenhagen)

Academic and Practitioners Chairs:

Tanu Mitra (University of Washington)
Giovanni Zagni (Pagella Politica / Facta)

Publicity Chairs:

Elena Kochkina (Queen Mary University of London, Alan Turing Institute)
Nevin Thompson (Hacks/Hackers)

Liaison Chairs – Academia, Industry, and Practitioner:

Tanu Mitra (University of Washington)
Giovanni Zagni (Pagella Politica / Facta)

Sponsorship Chairs:

Marzieh Saeidi (Facebook)
Jennifer Lee (Hacks/Hackers)

Technology and Website Chair:

Georgi Karadzhov (University of Cambridge)

Program Committee:

Oana Cocarascu (King's College London)
Emilio Ferrara (University of Southern California)
Fabiana Zollo (Ca' Foscari University of Venice)
Emiliano De Cristofaro (University College London)
Ioana Manolescu (Institut Polytechnique de Paris)
Preslav Nakov (Computing Research Institute, HBKU)
Gerhard Weikum (Max Planck Institute for Informatics)
Srijan Kumar (Georgia Institute of Technology)

Andreas Vlachos (University of Cambridge)
Ahmet Aker (University of Duisburg Essen)
Alberto Barrón-Cedeño (University of Bologna)
Gianluca Demartini (The University of Queensland)
Friedolin Merhout (University of Copenhagen)
Arkaitz Zubiaga (Queen Mary University of London)
Piotr Przybyła (Institute of Computer Science, Polish Academy of Sciences)
Kalina Bontcheva (The University of Sheffield)
Elena Kochkina (Queen Mary University)
David Corney (Full Fact)
James Thorne (University of Cambridge)
Naeemul Hassan (University of Maryland)
Kristen Johnson (Michigan State University)
Andreas Vlachos (University of Cambridge)
Dilek Hakkani-Tur (Google)
Harith Alani (The Open University)
Yevgeniy Golovchenko (University of Copenhagen)
Kalina Bontcheva (The University of Sheffield)
Jon Roozenbeek (University of Copenhagen)
Diana Maynard (The University of Sheffield)
Francesca Spezzano (Boise State University)
Denis Teyssou (Agence France-Press)
Saravanan Thirumuruganathan (QCRI)
Mohammed Saeed (EURECOM)
Sheikh Muhammad Sarwar (UMass)
Stefano Cresci (CNR)

Operations Manager:

Ahmed Medien (Hacks/Hackers)

Accepted Talks

Detecting Harm in Voice Communications
Mike Pappas

An Overview of Research on Knowledge Integrity in Wikimedia Projects
Diego Saez-Trumper and Pablo Aragon

Misinformation interventions are common, divisive, and poorly understood. How should they be designed
Claire Leibowicz, Emily Saltz and Soubhik Barari

Visualizing the Dimensions of Disinformation Campaigns
Amruta Deshpande and Justin Hendrix

Alternative Monetization on YouTube
Yiqing Hua and Manoel Horta Ribeiro

Online misinformation is linked to COVID-19 vaccination hesitancy and refusal
Francesco Pierri, Brea Perry, Matthew R. Deverna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer and John Bryden

Table of Contents

Conference Program

e-Game of FAME: Automatic Detection of FAke MEmes
Bahruz Jabiyev, Jeremiah Onaolapo, Gianluca Stringhini and Engin Kirda

People Expect Joint Accountability for Online Misinformation
Gabriel Lima, Jiyoung Han and Meeyoung Cha

E-BART: Jointly Predicting and Explaining Truthfulness
Erik Brand, Kevin Roitero, Michael Soprano and Gianluca Demartini

The Emergence of Deepfakes and its Societal Implications: A Systematic Review
Dilrukshi Gamage, Kazutoshi Sasahara and Jiayu Chen

Human-in-the-Loop Systems for Truthfulness: A Study of Human and Machine Confidence
Yunke Qu, Kevin Roitero, Stefano Mizzaro, Damiano Spina and Gianluca Demartini

Cross-lingual Rumour Stance Classification: a First Study with BERT and Machine Translation
Carolina Scarton and Yue Li

Game of FAME: Automatic Detection of FAke MEmes

Bahruz Jabiyev
Northeastern University
Boston, MA
bahruz@ccs.neu.edu

Jeremiah Onaolapo
University of Vermont
Burlington, VT
jeremiah.onaolapo@uvm.edu

Gianluca Stringhini
Boston University
Boston, MA
gian@bu.edu

Engin Kirda
Northeastern University
Boston, MA
ek@ccs.neu.edu

Abstract

Memes nowadays are ubiquitous on the Web and play a major role in disinformation campaigns. It is therefore not enough to tackle only the problem of textual disinformation. The research community must also develop new techniques to address the problem of malicious memes (fake memes) that contain misattributed or fabricated quotes, for instance, in online smear campaigns that target politicians and celebrities. To address this problem, we develop a system to automatically detect fake memes; our approach leverages optical character recognition, natural language processing, image processing, and machine learning techniques to carry out this task. Our implementation, a system named FAME, relies on various features to detect visual memes that contain fake or misattributed quotes. FAME classifies memes with 84% true positive rate and 14% false positive rate. It can be used for early detection of meme-based disinformation campaigns, for instance, if deployed on online social networks or messaging applications. To the best of our knowledge, FAME is the first automatic fact-checking tool for memes.

1 Introduction

Recent developments have demonstrated a relatively new mode of information warfare: attempts were allegedly made to influence the 2016 US presidential elections, among others, via coordinated disinformation campaigns on the Internet. Hordes of fake news articles (Allcott and Gentzkow, 2017), politically-motivated images (Zannettou et al., 2019a), and targeted ads (Wakefield, 2018) on online social networks (OSNs) played major roles in the push to sway public opinion and manipulate elections.

Images play an interesting role in information warfare: Zannettou et al. (Zannettou et al., 2019a) reported that state-sponsored actors "do not only use textual content, but also take advantage of the expressive power of images and pictures." Memes—a popular Internet vehicle of information that often involves attention-grabbing images—have also been co-opted by such actors; they create and disseminate memes with biased political messages, usually via OSNs and other online communities (Zannettou et al., 2018). Figure 1 shows an example of a politically-inclined meme.

Figure 1: A quote attributed to former British prime minister, Winston Churchill, which was determined to be a misattribution by Snopes, a fact-checking organization.

Previous work has studied fake news on the Internet and developed techniques to automatically detect fake news (Zhou and Zafarani, 2018). Despite these efforts, fake news is an ongoing problem and deserves further attention: early detection of fake news is one of the open challenges identified by Zhou and Zafarani (Zhou and Zafarani, 2018). On a related note, image-based disinformation, for instance via politically-charged memes, is an understudied field. Hence, we focus our attention on this research gap: we aim to detect image-rich disinformation content in order to mitigate disinformation campaigns on the Internet.

In this paper, we address the problem of fake memes—these contain messages, fabricated or otherwise, falsely attributed to specific individuals. Such memes could be deployed against political opponents during smear campaigns, for in-

Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 1–11,
Virtual, October 16-17, 2020.

stance. Our approach leverages Optical Character Recognition (OCR), Natural Language Processing (NLP), image processing, and machine learning techniques to detect memes that contain fake or falsely-attributed content, as previously described. Our implementation, a system named FAME (a contraction of "FAke MEmes"), relies on several information feeds to carry out its task: reputable news sources, quotation websites, verified social media accounts, and public government websites. FAME achieves 84% true positive rate and 14% false positive rate.

There is a caveat associated with FAME's false positive rate: meme classification is a hard problem that involves many complex interconnected tasks, including OCR, face recognition, and NLP, each with its own limitations. In Sections 6 and 7, we discuss how these limitations contribute to false positives. We also suggest potential ways to improve future instantiations of FAME; a key recommendation is to use high-performance proprietary OCR tools rather than free OCR tools (we used a free one in this work). Similarly, using proprietary tools for the other components—for instance, NLP and face recognition—would drastically reduce FAME's false positive rate.

FAME can be deployed by various digital platforms to stem the flow of meme-based disinformation campaigns. FAME's end goal is to make the Internet safer for the general public. Our contributions are as follows.

- We identify features for the classification of fake memes; these include reputable news sources, quotation websites, verified social media accounts, and public government websites.
- We develop a novel approach for the automatic detection of fake quotes and falsely-attributed quotes in images.
- We make the source code of FAME available to the public so it can be deployed by OSNs, messaging apps, and other platforms to stem image-based disinformation campaigns. The code is publicly available on the authors' websites.
- We evaluate FAME's performance and discuss potential ways to improve it.
- We create a labeled dataset (FAME dataset) which contains 1000 fake and real quote memes, for future research into understanding and mitigating disinformation campaigns on the Internet. The dataset is publicly available on the authors' websites.

2 Background and Related Work

To help the reader understand the remainder of this paper, this section presents the three main themes that comprise the foundation of our work: fake news, memes, and fact checking.

2.1 Fake News

Fake news, according to Allcott and Gentzkow (Allcott and Gentzkow, 2017), comprises "news articles that are intentionally and verifiably false, and could mislead readers." Although fake news is not a new phenomenon, (Soll, 2016) it again came into the public spotlight during the 2016 US presidential elections, in which political actors allegedly attempted to manipulate public opinion via fake news and other methods. Unfortunately, current efforts to stem the spread of fake news have not yet recorded much success (Lee, 2016). Prior work on the detection of fake news includes (Tacchini et al., 2017; Tschiatschek et al., 2018; Zhou et al., 2015; Jin et al., 2016; Volkova et al., 2017; Liu and Wu, 2018; Ruchansky et al., 2017; Wang et al., 2018; Yang et al., 2018). Other studies on the propagation of false or malicious information include (Zannettou et al., 2019b; Zhou and Zafarani, 2018; Zannettou et al., 2017; Hine et al., 2017; Zhang et al., 2018).

2.2 Memes

According to Richard Dawkins (Dawkins, 1976), a meme—analogous to a gene—is an idea or unit of culture that is replicated and transmitted among people. Internet memes, often comprising catchy images and text, are transmitted via numerous online communities and social networks, sometimes for comedic effect, and other times with malicious intent. Internet memes are ubiquitous nowadays, and successful memes spread rapidly through various online communities (Bauckhage, 2011; Zannettou et al., 2018). Hence, memes are attractive to malicious actors who intend to carry out disinformation campaigns (Zannettou et al., 2019a). Memes often originate from fringe online communities (Zannettou et al., 2018) and then spread to the rest of the Web. For instance, 4chan, an online

message board, is reportedly the source of many popular politically-charged memes (Hine et al., 2017).

2.3 Fact Checking

Fact checking is one of the approaches that have been deployed to tackle fake news. At its core, fact checking involves comparing news content to well-established facts to ascertain if the news content under test is true or not. It can be carried out manually (by credible domain experts) or automatically (using information-processing software). Manual fact-checking, although often accurate, does not scale well, given the sheer volume of content that online communities produce daily (Zhou and Zafarani, 2018; Zannettou et al., 2018). Existing fact-checking services include Snopes, PolitiFact and FullFact which provide manual fact-checking services. Memechecker.net is another fact-checking service which focuses its efforts only on memes, while listing much fewer – less than a dozen – fact-check reports than the mentioned fact-checker organizations. Since memes play a vital role in disinformation campaigns as discussed in Section 2.2, our work aims to provide a scalable solution to the problem of fact-checking memes. In other words, we propose an automatic meme fact checker to help increase the scale of fact checking and minimize the potential psychological harm that human fact checkers encounter during their work.

3 Problem Statement

Fake quote memes comprise images which contain fake quotes, usually attributed to well-known people. They exist mainly in three forms: memes with fabricated quotes (made up), slightly-modified real quotes (slight change in the text, usually significant change in the meaning), and misattributed quotes (real quotes that actually originated from someone else not present in the meme). We focus on memes that contain fabricated or misattributed quotes.

We only address quote memes which attribute exactly one quote to one person, for technical reasons. We exclude memes that contain several persons or multiple quotes. Consider the worst-case scenario: a quote meme that contains several persons and multiple quotes. Limitations in image processing and OCR techniques prevent us from successfully matching such quotes to individual persons on the meme. Hence, as earlier mentioned, we focus on simple quote memes: one person, one quote. Figure 1 shows such an example.

Purveyors of false information via memes do not always include quotes in their memes. Sometimes, they opt for doctored images without text captions. For example, they might take a picture from a gruesome murder scene and edit it to replace the victim's face with the face of their target (say, a politician or celebrity). Such memes are out of scope in this work; they require a different approach than ours.

Our work aims to protect vulnerable online communities and digital platforms that double as sources of information, from certain types of image-based disinformation campaigns. Scenarios in which our work will be directly applicable include the following: (1) a journalist may be targeted with malicious fake quote memes in the course of reporting sensitive events, and (2) OSNs, which double as news sources for millions of people, may be contaminated with fake quote memes to defame high-profile individuals, for instance, during elections.

4 An Overview of Our Approach

Figure 2 illustrates an overview of FAME's steps in classifying meme quotes. Next, we discuss each step in detail.

4.1 Extracting Meme Text

To extract text from the input meme, we use OCR, a technique commonly used to extract text content from images and Portable Document Format (PDF) files. Factors that may affect the quality of OCR include color contrast between the text and background, font family of the text, and the amount of distortion in the text segment of the image. In Section 6, we discuss the performance of text extraction and how it affects the meme classification process.

4.2 Identifying the Subject

To identify the subject (person) to whom the quote is attributed, we use two techniques: recognition of the person's *face* from the meme and recognition of the person's *name* from the caption of the meme. To identify the person on a meme, we first perform face recognition on the meme. However, quote memes do not always display a face; instead, some include the person's name in text only. Also,

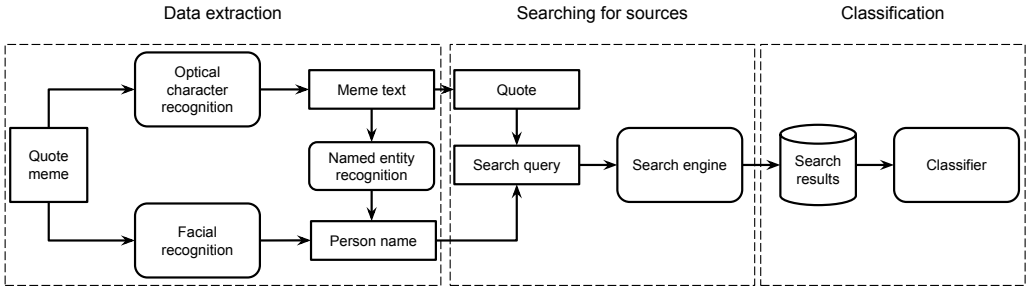


Figure 2: An overview of our meme classification pipeline.

face recognition sometimes fails. In such cases, we attempt to deduce the name of the subject from the text extracted from the meme (using OCR). For this, we leverage Named-Entity Recognition (NER), a technique for identifying names of various entities (for instance, people, organizations, and places) within a body of text.

4.3 Searching for Sources

To search for sources, we first have to obtain the meme quote in question—it has to be included in the search query. To find the quote within the text extracted using OCR, we retrieve the text segment that is enclosed in double quotation marks. However, sometimes we fail to extract the quote for two reasons: either OCR fails to recognize double quotes or the meme text does not contain double quotes. Our observations reveal that non-quote text, which sometimes appears on memes alongside quote text, might prevent search engines from retrieving sources for the quote when included in the search query. Hence, we construct search queries in two distinct ways, depending on our ability to find the quote in the OCR-extracted meme text. We discuss them next.

Success during quote extraction. When the OCR-extracted text contains a pair of double quotation marks with a body of text between them, we assume that body of text is the quote. We include the extracted quote in our search query as it is.

Failure during quote extraction. If OCR fails to recognize double quotation marks in the text or the text actually does not contain quotation marks, we split the text into sentences—knowing that at least one of them belongs to the quote segment—and construct a separate search query for each of them. Search queries which contain a sentence from the quote text will return sources for the quote (if they exist), while search queries which contain content from non-quote text will not return such sources. Additionally, we construct yet another search query using the whole OCR-extracted text.

Constructing search queries. We take the following steps to construct a search query whether we succeed or fail to identify the quote segment. First, we remove misspelled words to avoid confusing the search engine. Second, we trim the search query to its first n words—Appendix A discusses how we arrived at this—having observed that the first n words were sufficient for the search engine to recognize the quote. Third, we add the name of the subject (person on the meme) to the beginning of the search query, because it helps the search engine to return more relevant results. Finally, we submit search queries to the search engine—or only one search query if the quote segment has been successfully retrieved, as discussed previously.

Outcome. We combine the search results to create a *pool of retrieved search results* after removing duplicate results.

4.4 Identifying Relevant Search Results

Not all search results returned by the search engine will be relevant. To identify the relevant ones from the pool of retrieved search results, we test two conditions: quote condition (to ensure that a search result page includes the quote) and name condition (to ensure that a search result page contains the name of the person on the meme).

Outcome. We create a *pool of relevant search results* by applying both conditions to the pool of retrieved search results.

4.5 Classification

In this section, we discuss several features of the pool of relevant search results that serve as inputs for the meme classification task. We enter those features into a trained machine learning model to

compute the probability of the input quote meme being fake or real.

Highly-trusted news sources. These comprise a selection of news sources that have established a strong reputation, over decades, of reliable and accurate reporting, and by having high standards of reporting. We refer to them as highly-trusted sources throughout this paper. We compute the number of these sources from the pool of relevant search results.

Legitimate news sources. There is a large number of online news sources which do not necessarily carry out in-depth investigation and reporting as highly-trusted sources do, yet are known to be reliable. We call these sources legitimate news sources. We count the number of legitimate news sources in the pool of relevant search results.

Quotation websites. Compared to well-known living people today, it is harder to find quotes of well-known historical figures in news sources. Hence, we use quotation websites as sources when searching for a quote. We count the number of quotation websites in the pool of relevant search results.

Government websites. Government websites are usually reliable sources of quoted information, especially from politicians, who also happen to be common targets of fake quotes. Hence, we count the number of government websites in the pool of relevant search results.

Verified social media accounts. Finally, we check for the existence of verified social media accounts of the subject in the pool of relevant search results.

5 Prototype Implementation

In this section, we present our implementation of the quote meme classifier which we call FAME (a contraction of "FAke MEmes"). It is based on our general approach to the quote meme classification task, as discussed in Section 4.

5.1 Extracting Meme Text

To extract text from a meme, we use a free OCR API called *OCR.space*.[1] This API receives image information either via a URL that points to an image, or the image itself as a base64-encoded string, and returns the extracted text. We discuss the performance of this API in Section 6, with emphasis on how it affects the performance of our classification model. If the extracted text from a meme does not contain a sentence with at least three words, we discard that meme.

5.2 Identifying the Subject

To identify the person on a quote meme, we use two techniques: face recognition and name recognition, as mentioned in Section 4. For face recognition, we rely on the image search function that the Bing search engine provides. To this end, we craft an HTTP request, include the URL of the quote meme in it, and carry out an image search on Bing. We then retrieve the name of the person on the meme from the "Looks like" section of the resulting HTTP response. We discuss details of the performance of Bing image search in Section 6.

If face recognition fails, we run person name recognition, otherwise known as Named-Entity Recognition (NER), on OCR-extracted meme text. To this end, we use *CoreNLP*, a Natural Language Processing library, to implement NER. We discuss its performance in Section 6.

5.3 Searching for Sources

Preprocessing. In Section 4, we explained the process of constructing search queries from meme text. This process requires the removal of misspelled words as a preprocessing operation; we use a library called *pyenchant* to achieve this.

Search engine. We chose DuckDuckGo to search for sources. Arguably, using another search engine such as Google might result in better performance. However, Google blocks scripted requests and would not allow us to run as many queries as required; sometimes, in experiments, we made about 2000 requests within a few hours. DuckDuckGo sources results[2] from several partners including Bing and Yahoo. In Section 6, we show that situations in which DuckDuckGo is unable to retrieve sources that Google can fetch, are very few. To query DuckDuckGo, we craft and send HTTP requests, and use only the first two pages of search results to find sources.

5.4 Identifying Relevant Search Results

We use a Python library called *edit_distance* to implement the quote condition (see Section 4.4), which looks for the quote of interest within the page of a search result. To carry out the longest

[1] https://ocr.space/ocrapi

[2] https://help.duckduckgo.com/duckduckgo-help-pages/results/sources/

common subsequence task, we use the value *highest_match_action* for the parameter of *action_function*. Once the longest common subsequence is found, we check it against the threshold value *0.3*, which is the ratio of the length of the longest common subsequence to the length of the quote (or meme text if the quote cannot be extracted). If the length ratio is above that threshold, we add the corresponding search result to the pool of relevant search results. We discuss how we chose this threshold value in Appendix A.

To check the name condition, we use a simple regular expression to search for the identified person's name in a search result page.

5.5 Classification

To implement the FAME classifier, we use *scikit-learn*, a Python library. As we show in Section 6, Support Vector Machine (SVM) with rbf kernel yields the best results in the quote meme classification task. To extract information regarding the features mentioned in Section 4, we do the following.

Highly-trusted news sources. To implement this feature, we create a list of news sources based on the results of two separate public surveys conducted by Pew Research Center (Center, 2014) and Reynolds Journalism Institute (Kearney, 2017). This list comprises about 30 different news sources (see Appendix C). We use it to identify the number of highly-trusted sources in the pool of relevant search results, by counting how many domain names of search results match domain names in the list of highly-trusted news sources.

Legitimate news sources. Similarly, we check each search result in the pool of relevant search results against a list of legitimate news sources. This list comprises the Alexa Top 500 newspaper websites in the United States, with a slight modification; we remove highly-trusted news sources to eliminate repetition in counting.

Quotation websites. We check each search result in the pool of relevant search results against a list of quotation websites. For this list, we use Alexa Top 500 Quotation websites; it contains about 130 websites.

Government websites. To identify government websites in the pool of relevant search results, we specifically search for US government websites and use a simple regular expression which checks if the domain name of a search result ends with ".gov" or not.

Verified social media accounts. After we identify search results that point to Twitter or Facebook profiles, we carry out a scripted HTTP request to identify if they are verified or not, and also obtain the full name on the profile. If they are verified, we then check the name of the person of interest (which we extract from the input meme) against the full name on the page.

6 Evaluation

In this section, we discuss our ground truth dataset and evaluate the performance of our classification model. We also discuss the performance of specific components of FAME.

6.1 Ground Truth Dataset

We evaluate our system on a quote meme dataset which we collected ourselves, called the *FAME dataset*. It contains 1000 quote memes in total: 379 fake memes and 621 real memes.

Collection. First, we identified 20 well-known individuals (see Appendix D) who are commonly targeted by fake quote memes, by analyzing the fact-check history of three main fact-checking organizations: Snopes, FactCheck, and PolitiFact. Next, we used the DuckDuckGo search engine to collect quote memes for each of them by entering "$\{person's\_name\}$ *quote memes*" in the search field. We avoided duplicate quotes across memes during the collection process.

Labeling. To complete the ground truth dataset, we needed binary labels for the memes: "real" or "fake." We followed a set of guidelines to label the memes. First, we searched for the quote on a search engine and examined the search results which contained the quote. We examined the domain names of those search results with the help of a browser plugin that we implemented specifically for this purpose. The plugin colorizes search results of interest, with unique colors, depending on their type: "highly-trusted news source," "legitimate news source," "quotation website," "government website", or "verified social media account" (as discussed in Section 5).

We labeled memes as "real" if they met either of these conditions: (1) they had at least one search result that published the quote and was a highly-trusted news source, government website, or verified social media account that belonged to

Table 1: Performance metrics of the classification model.

Metric	*Performance*
Accuracy	85%
True positive rate (recall)	84%
True negative rate	86%
Precision	79%
F1 score	81%
False positive rate	14%
False negative rate	16%

the identified person, or (2) they had at least two search results that published the quote and both of them were either a legitimate news source or quotation website. On the other hand, we labeled memes as "fake" if they met both of these conditions: (1) no search result, of the previously discussed types, published the quote, and (2) it did not appear credible that the words on the meme were uttered or written by the identified person on the meme. The second condition is necessary because we acknowledge that the absence of reliable sources in search results does not conclusively indicate that the quote is fake; search engines sometimes fail to retrieve sources.

6.2 Classification Performance

Five-fold cross validation. To evaluate our model, we applied the cross-validation technique on the FAME dataset. It involves splitting the dataset into n parts. During each of n iterations, one part is left out for testing and the rest of the dataset is used for training. Overall metrics of the model can be evaluated by computing the average of metrics achieved during each iteration. We achieved the highest performance with SVM classifier with rbf kernel (see the performance of other classifiers in Appendix B). SVM with five-fold cross validation gives the results in Table 1; the FAME classifier achieves 84% true positive rate and 14% false positive rate. Besides these metrics, we also evaluate the time taken during the classification of a quote meme. When we run our system on a Docker container with 16 CPUs and Ubuntu installed on it, the average amount of time taken for classification is about 20 seconds, including data extraction and searching of a meme.

False positives. There are several reasons why our prototype model mistakenly classifies real quote memes as "fake." A common reason for false positives is OCR failure. When OCR drops or misspells a significant portion of the meme text, the subsequent search engine query fails to retrieve sources based on that text. This reason is responsible for one-third of the false positives. We further discuss the performance of the OCR component in Section 6.3.

Another reason is that the lists we use to categorize sources do not—and presumably cannot—include all reliable and legitimate sources of information. When our system cannot match the domain names of search results with its lists, it simply assumes a lack of sources for the quote. We owe another one-third of false positives to this reason.

False positives also arise as a result of failure of the search engine to retrieve sources. In those cases, the search engine becomes confused by either non-quote text or inadequate quality of extracted text, and therefore fails. This reason accounts for one-fifth of false positives. We also searched for those failed search queries on Google and found that in one-third of those cases, it managed to retrieve sources for the quote.

In very few cases, false positives arise as a result of mistakes in identification of the meme subject. Either face recognition or named-entity recognition may mistakenly identify some other person to be the subject. Such cases confuse the search engine and it fails to retrieve sources.

False negatives. Similarly, our system sometimes misclassifies fake quote memes as "real," for several reasons. The most common reason for false negatives is, for some fake quotes, search results contain common words just as many as to reach the threshold by chance. This happens especially when the quote is short and contains common English words. This accounts for one-fourth of false negatives.

Another common reason is that we fail to identify misattribution on some memes. Usually when a meme attributes Person A's words to Person B, sources for the quote will contain the name of Person A, not Person B. However, in some cases, sources contain both names: the real author and the falsely-attributed author. This accounts for one-fifth of false negatives.

Also, OCR sometimes drops words to an extent where only a few words from the quote remain. When we search for them, search results can easily contain some parts of those few words by chance. This in turn causes our system to reason that those search results are sources for the quote. This ac-

counts for one-sixth of false negatives.

Some sources contain fake quotes, not for publishing, but instead for debunking purposes. Our system cannot distinguish such search results and sees them as sources for the quote. One-sixth of false negatives stem from this reason.

In a few cases, a wrong segment of text is extracted as the quote. Some memes contain quotes in a way that such quotes have some segments enclosed in double quotation marks (nested quotes). Those parts tend to be short, and when searched, some search results happen to have some words in common, in the same order. This accounts for one-tenth of false negatives.

6.3 Performance of Specific Components

Text extraction API. We evaluate the performance of *OCR.Space* (a free OCR tool) on 621 real memes drawn from the FAME dataset. For 5% of real memes, search queries could not be formed because the extracted text did not contain a sentence with at least three words. In another 5% of real memes, the poor quality of extracted text caused the search engine to fail while retrieving sources for the quote; if text quality were good, the search engine would have succeeded in retrieving sources. *OCR.Space* extracted text well enough on 90% of real memes, which allowed the search engine to retrieve sources for them.

Face recognition. Bing image search correctly identified the person on 69% of all memes in our dataset. On less than 1% of all memes, it confused the person on the meme with someone else. It failed to give any result on 30% of memes.

Name recognition. *CoreNLP* NER identified the quote author's name from meme text on 57% of memes that failed face recognition—about 300 memes. It could not extract the author's name correctly from 3% of them. It could not extract any name from the text—either because the name did not exist or it was missed by NER—on 40% of them. A part of *CoreNLP*'s failure can be attributed to the failure of OCR earlier in the pipeline.

7 Concluding Discussion

We focused on a non-trivial problem and developed an approach to detect memes which contain misattributed or fabricated quotes. As we have demonstrated, meme classification is a hard task that involves many interconnected components, each with its own limitations. In Section 6, we addressed those limitations in detail, and discussed how they contributed to FAME's false positive rate. Despite this, FAME's performance shows that our approach can be reliably adopted in practice. Its performance would be even better if we had access to a proprietary OCR tool, rather than the free one we used, and had extensive lists of reliable sources. This also applies to other components of FAME, including NLP and image recognition modules; proprietary tools would perform better and boost FAME's overall performance.

Search engines play a central role in our work: we used a search engine to query the sources of quotes and examine those sources for their reliability. We also searched for the names of people in sources to ensure that the quotes had not been misattributed. Search engines are neither perfect nor the only available tools; there are many other valuable resources and databases, some of which grant free access, while others charge fees. Occasionally during the labeling process, we could not find sources for a quote, using a search engine, and could not label it as "fake" because it did not seem unreasonable that the purported author said or wrote it. However, search engines give free and quick access to large numbers of online resources with a quick search; they are therefore commonly used and highly recommended by fact-checking organizations, for instance, AFP (AFP, 2011), FactCheck (Jackson, 2008) and PolitiFact (Holan, 2014). Our work offers a framework for future research that might use other resources and databases for the identification of fake and real quotes.

Our work offers significant benefits to fact-checking organizations that rely on manual fact-checking processes by experts and cannot handle large numbers of quote memes daily. Our system will help to scale up their work. To further improve their output, they can also compile their own lists of sources that they rely on, instead of using the ones that we compiled for the FAME prototype. In addition to the time-related benefits of scaling up, our approach will also help to minimize potentially harmful content that human fact checkers will be exposed to, which will in turn reduce mental trauma, as mentioned in Section 2.3.

Finally, our approach can also be used by messaging apps and digital platforms that host quote memes. They can leverage our work to automati-

cally detect fake and real quote memes—in a reasonable amount of processing time—uploaded to their platforms. This will help in *early* stemming of disinformation campaigns, towards making the Internet safer for everyone.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported by National Science Foundation under award numbers of CNS-1942610, CNS-2114407, and CNS-2127232.

References

AFP. 2011. Fact-checking : how we work. https://factcheck.afp.com/fact-checking-how-we-work.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Christian Bauckhage. 2011. Insights into internet memes. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.

Pew Research Center. 2014. Political polarization & media habits.

Richard Dawkins. 1976. *The selfish gene*. Oxford University Press.

Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Eleventh International AAAI Conference on Web and Social Media*.

Angie Drobnic Holan. 2014. 7 steps to better fact-checking. https://www.politifact.com/article/2014/aug/20/7-steps-better-fact-checking/.

Brooks Jackson. 2008. Obama Quote Rumors. https://www.factcheck.org/2008/08/obama-quote-rumors/.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Michael W. Kearney. 2017. Trusting news project report.

Dave Lee. 2016. Facebook's fake news crisis deepens. https://www.bbc.com/news/technology-37983571.

Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Jacob Soll. 2016. The Long and Brutal History of Fake News. https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/.

Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Jane Wakefield. 2018. Cambridge Analytica: Can targeted online ads really change a voter's behaviour? https://www.bbc.com/news/technology-43489408.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2019a. Characterizing the use of images by state-sponsored troll accounts on twitter. *arXiv preprint arXiv:1901.05997*.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *ACM Internet Measurement Conference (IMC)*.

Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019b. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data and Information Quality*, 11(3).

Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B Gouza. 2018. Fake news detection with deep diffusive network model. *arXiv preprint arXiv:1805.08751*.

Xing Zhou, Juan Cao, Zhiwei Jin, Fei Xie, Yu Su, Dafeng Chu, Xuehui Cao, and Junqiang Zhang. 2015. Real-time news certification system on sina weibo. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.

A Setting Thresholds and Parameters

Length ratio threshold. This threshold determines if a search result meets the quote condition; it checks if a search result published the quote. We tried multiple values—between 0.25 and 0.5—for this threshold and picked 0.3 which ensures the best balance between precision and recall.[3] It also yields high accuracy. Table 2 shows changes in performance metrics relative to the performance achieved by the baseline value (0.3).

Table 2: Changes in performance metrics relative to the performance achieved by the baseline length ratio threshold (LRT). The baseline LRT is in boldface.

LRT	*Accuracy*	*Recall*	*Precision*	*F1 score*
0.25	-3%	-12%	0%	-6%
0.30	**85%**	**84%**	**79%**	**81%**
0.35	-1%	+3%	-3%	0%
0.40	-1%	+5%	-5%	0%
0.45	-2%	+8%	-7%	-1%
0.50	-4%	+8%	-10%	-2%

Optimal query length. During the labeling process, we observed that it suffices to use the first 20 words (approximately) of the quote to search for its sources. Therefore, we tried multiple values around that length—15, 20, and 25 words—and decided on 20 words: using 20 words resulted in 1% better recall than 15 words, and 1% better recall and precision than 25 words.

Window size. This parameter comes into play when we want to limit the distance between found words in search results to ensure that those found words are adjacent. Initially, we set the window size to twice the length of the quote. However, we achieved better performance—2% better recall—when we set it equal to the length of the quote.

[3] Recall is also known as the true positive rate.

B Comparison with other classifiers

We also compare our classification algorithm, SVM with rbf kernel, with other algorithms. As Table 3 shows, other classification algorithms also performed well and some surpassed our classification algorithm in some metrics. Nonetheless, we chose SVM with rbf kernel to create a balance between precision and recall, and at the same time achieve the highest accuracy and F1 score.

Table 3: Performance of other classification algorithms compared to our choice. Our choice is in boldface.

	Accuracy	*Recall*	*Precision*	*F1*
SVM (rbf)	**85%**	**84%**	**79%**	**81%**
SVM (linear)	82%	89%	71%	79%
Random Forest	84%	84%	77%	80%
KNN-3	82%	88%	71%	79%
Adaboost	84%	85%	76%	81%

C Highly-trusted Sources

- ABC News
- Associated Press
- BBC
- Bloomberg
- CBS News
- CNN
- Dallas News
- Fox News
- Google News
- Los Angeles Times
- MSNBC
- NBC News
- NPR
- PBS

- Politico
- Reuters
- The Atlantic
- The Denver Post
- The Economist
- The Guardian
- The Kansas City Star
- The New York Times
- The New Yorker
- The Seattle Times
- The Wall Street Journal
- The Washington Post
- TheBlaze
- Time
- USA Today
- Yahoo News

D Well-known Individuals

- Alexandria Ocasio-Cortez
- Barack Obama
- Ben Carson
- Bernie Sanders
- Bill Murray
- Donald Trump
- Elizabeth Warren
- Hillary Clinton
- Ilhan Omar
- Kurt Russell
- Melania Trump
- Michele Bachmann
- Michelle Obama
- Nancy Pelosi
- Ronald Reagan
- Ruth Bader Ginsburg
- Sarah Palin
- Stacey Abrams
- Ted Cruz
- Winston Churchill

People Expect Joint Accountability for Online Misinformation

Gabriel Lima[1,2], **Jiyoung Han**[3], **Meeyoung Cha**[2,1]
[1] School of Computing, KAIST, Daejeon, South Korea
[2] Data Science Group, Institute for Basic Science (IBS), Daejeon, South Korea
[3] Moon Soul Graduate School of Future Strategy, KAIST, Daejeon, South Korea
{gabriel.lima, jiyoung.han}@kaist.ac.kr, mcha@ibs.re.kr

Abstract

Identifying who should take responsibility for online misinformation is critical for mitigating its detrimental effects on society. This research offers a multi-faceted picture of the public's perception on who is responsible for false information online separately for 1) creating, 2) disseminating, and 3) failing to prevent it. Our study (N=496) shows that the responsible entities differ across distinct aspects of online misinformation. For instance, people and interest groups are associated with creating falsehoods, whereas social media platforms are predominantly seen as accountable for failing to prevent them. We discuss several implications, including the public demand for accountable social and news platforms and the importance of joint accountability in the fight against online misinformation.

1 Introduction

Who should be blamed for creating and disseminating misinformation online or failing to prevent falsehoods from reaching a wide audience? While much research has been devoted to understanding how misinformation travels (Vosoughi et al., 2018; Kwon et al., 2013; Shao et al., 2018), the question of whom the general public views as the main actors in its creation, dissemination, and prevention remains open. Answering this question is imperative to designing policies and interventions that can combat misinformation, such as regulation (Cha et al., 2020) and online interventions (Pennycook et al., 2021).

Scholars, news reporters, and other stakeholders often discuss whom to blame for the uncontrolled spread of misinformation online and thus should take the lead in the fight against it. Following the influence of social media in the 2016 US election, news media quickly turned to social media platforms, particularly Facebook, for accountability (The Atlantic, 2017). Following the election, Mark Zuckerberg, Facebook's CEO, defended the platform's stance in not taking a proactive role in content moderation by stating that Facebook did not wish to be "arbiters of truth" (The Walt Street Journal, 2016).

After widespread backlash from the public and mainstream media, social media platforms decided to take a more active role in the 2020 US election by flagging misleading posts and removing false conspiracy theories (The New York Times, 2020). Nevertheless, some argue these efforts are not enough to "save democracy" (The Washington Post, 2020b).

Another perspective has instead underscored journalists' and news institutions' role in disseminating misinformation. First, mainstream media might contribute to the dissemination of falsehoods through their debunking efforts, although not intentionally as other nefarious actors (Tsfati et al., 2020). People might be exposed to false information due to mainstream media unnecessarily correcting falsehoods that would otherwise only reach a small number of citizens.

Second, journalists might disseminate misinformation due to media manipulation. Interest groups, such as conspiracy theorists and trolls, have developed techniques to increase their visibility by targeting news media sources to disseminate their content (Donovan and Friedberg, 2019; Marwick and Lewis, 2017). For instance, these groups can coordinate actions that force specific topics into the public discourse that journalists may not be able to ignore. Journalists are aware of such propaganda; however, they report barriers in delivering accurate information, such as technical difficulties in obtaining data from social media and the power relations between them and online platforms (McClure Haughey et al., 2020; Balod and Hameleers, 2019).

Instead of focusing on a specific actor's re-

Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 12–17,
Virtual, October 16-17, 2020.

sponsibility, Graves and Wells (2019) have argued that political elites, news media, and citizens all have their roles in establishing "factual accountability" through a collaborative effort. Public discourse should not be limited to specific entities that are expected to prevent misinformation online. Rather, it should embrace a joint responsibility undertaking in which various actors are held jointly accountable for online information.

These normative approaches to responsibility often neglect how online users, and more generally the general public, perceive all stakeholders' roles in the spread of misinformation. Descriptive analyses of this question have been limited. A 2016 Pew Research Center study found that US adults considered the general public, politicians, and social media platforms similarly responsible for not preventing fake news from gaining attention (Barthel et al., 2016). In contrast, a later study from 2018 indicated that although politicians and activist groups are blamed for creating false claims, news media platforms are expected to take the lead in reducing the spread of falsehoods (Mitchell et al., 2019). Albeit helpful, these results are hard to compare as a whole. Our study inquired who is to blame for online misinformation with respect to its creation, dissemination, and the failure for prevention—this multi-faceted view of misinformation has not been studied systematically.

2 Methods

We conducted an online survey to answer this question. After agreeing to the research terms, participants were shown a short definition of online misinformation. Participants were then asked whom they consider responsible for the three aforementioned aspects of online misinformation: creation, dissemination, and prevention. The study ended with a set of demographic questions.

2.1 Respondents

We recruited 500 participants through Prolific during May 3rd-4th, 2021. Prolific is a crowdsourcing platform for recruiting subjects for social and economic experiments (Palan and Schitter, 2018). Our study was restricted to US residents who had previously completed at least 100 tasks with a minimum approval rate of 95%. Respondents were compensated US$1.43 for their participation. Four participants were removed for failing a simple attention check question that had instructed them to choose a specific answer. The final dataset analyzed was composed of 496 respondents. Women comprised 38.7% of the sample, and nearly half of participants were younger than 35 years old (51.8%; M=37.0, SD=12.4). A majority of respondents had received a Bachelor's degree (63.7%) and identified themselves as Caucasian (63.7%). African Americans represented 14.9%, whereas Asians and Hispanics comprised 12.9% and 5.2%, respectively. Participants were Democrats (41.3%) or Republicans (32.7%).

2.2 Measures

Participants were asked which entities they found responsible for 1) creating, 2) disseminating, and 3) not preventing the dissemination of false information online. They were shown a list of entities: the general public, social media users, social media companies, people with vested interests (i.e., interest groups), conspiracy theorists, news media, politicians, national institutions (that showed the government and the FCC as examples), and foreign institutions (showing foreign governments and actors). This list was compiled from a pilot study we do not report in this paper. Each participant chose as many entities as they wished for each aspect of online misinformation. Participants were also allowed to write down any other responsible entity in free text form. Each aspect was presented separately and in random order; entities' presentation order was randomized between subjects.

3 Results

We employed chi-square tests to identify whether participants held different actors responsible for distinct aspects of online misinformation. Participants' attribution of responsibility differed between creation, dissemination, and prevention ($\chi^2(16)$=338.53, p<.001, Cramer's V=0.16; see Figure 1). Social media users, conspiracy theorists, and interest groups were deemed the most responsible for creating online misinformation, followed by politicians and news media outlets. Foreign institutions and the general public came next and were followed by social media platforms as actors moderately responsible for creating false claims. Local institutions were not perceived to play a major role in the creation of false information online.

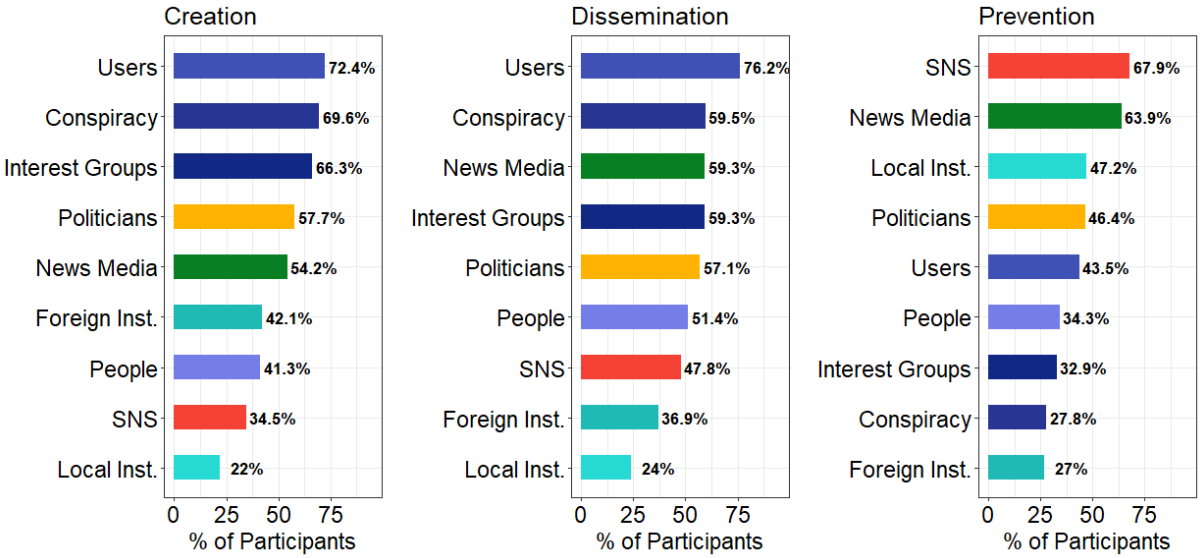


Figure 1: *"Which party or whom do you find responsible for creating, disseminating, and not preventing the dissemination of false information?"* The percentage of participants who picked each entity to the question of who is responsible for creating (top), disseminating (middle), and not preventing (bottom) misinformation online. Participants were shown all entities in random order and could select as many as they wished.

Social media users were the actors most blamed for disseminating false information. A series of entities were deemed moderately responsible, indicating a more homogeneous distribution of blame. People and social media platforms were deemed moderately responsible, followed by foreign and local institutions.

Social media platforms and news media outlets were perceived as the actors most responsible for the failure to prevent misinformation. Local institutions, social media users, and politicians were blamed to a moderate level. People, interest groups, conspiracy theorists, and foreign institutions were deemed the least responsible actors concerning prevention.

Only 28 out of 496 (5.64%) participants provided any responses in free text. Most referred to politics-related actors (e.g., "Liberals," "Republicans") as if participants meant to highlight which side of the political spectrum they blamed. However, we did not observe any substantial difference in the entities held responsible across political partisanship (p-values for all χ^2-tests greater than .05, Cramer's V smaller than 0.06). This finding suggests that people blame the same entities (e.g., news media) but may focus on those they disagree with politically.

Our data showed that people blame different entities for distinct aspects of online misinformation. Social media users and those with vested interests in creating misinformation were blamed the most for creating falsehoods, followed by politicians and news media outlets. These same actors were also deemed homogeneously responsible for the dissemination of misinformation. In contrast, social media platforms and news media outlets were expected to prevent false information already out there, followed by local governments, politicians, and users. We discuss the implications of these findings below.

4 Implications

4.1 Public Demand for Accountability

Social media platforms have been widely accused of spreading misinformation online (The Atlantic, 2017; The Washington Post, 2020a). Our results concerning prevention concur with this perspective. Current social media platforms' efforts addressing political misinformation during and following the 2020 US election (The New York Times, 2020; CNN, 2020) agree with this stance. Nevertheless, we highlight that similar efforts should be extended to other countries (e.g., Brazil (New York Times, 2018), the EU (BBC, 2019a)) and events (e.g., the COVID-19 vaccine rollout (The Washington Post, 2019)).

Another important public demand for accountability was directed at news media. News organizations were highly blamed for creating, dissemi-

nating, and failing to prevent misinformation, indicating that although news media express their aversion to it and often blame social media platforms, people have a different perspective. Hence, news media outlets should be aware of their expected role in all aspects of misinformation to mitigate its ill effects.

4.2 Government's Non-Partisanship Role

Participants believed local governments should rise to the challenge and work towards preventing online misinformation. Hence, regulation targeting misinformation could be perceived as beneficial by the public, particularly if it leads to accountability for social media platforms and news media, as discussed above. Participants also moderately blamed foreign actors for creating and disseminating misinformation. This could have been caused by the participants' context as foreign interference with US democratic institutions has been widely reported in recent years (The New York Times, 2021). Nevertheless, we highlight that international collaborations to combat and prevent falsehoods are important, as online misinformation is a global problem that does not respect territorial boundaries.

4.3 Fighting Interest Groups

Participants deemed interest groups responsible for creating and disseminating misinformation online. Unfortunately, this public expectation goes against existing cases where interest groups were found to play crucial roles and were not held accountable. For instance, various reports indicate Macedonia youth's role in creating and spreading political falsehoods during the 2016 US election (BBC, 2019b); those who took part in it justified their actions with financial reasons.

Financial motivations may be an important driver of misinformation. Misinformation should thus be fought against at its financial core. One may envision reforming Ad-based revenue models. Those who create and disseminate misinformation can substantially profit through Ad-based revenue models (Funke et al., 2019; Braun and Eklund, 2019). These paradigms should not promote false information but foster trustworthy news sources better. Similar efforts addressing misinformation's financial components could be crucial to mitigating intentional online misinformation.

5 Concluding Remarks

5.1 Expectation of Joint Responsibility

The findings in this paper shed light on how misinformation can be dealt with in an online environment. One of the key findings is that participants held different actors responsible for distinct aspects of online misinformation, highlighting that any intervention to address or regulate it should be crafted with specific actors and objectives in mind. For instance, addressing only certain aspects of it, e.g., by regulating social media platforms, might help prevent misinformation but not weaken its creation. Focusing on specific entities might not holistically deal with the complex issue of online misinformation.

It is worth noting that no entity was deemed solely responsible for the holistic problem of online misinformation, i.e., participants did not single out an individual or organization that should be held to account. This trend is particularly prominent when considering misinformation dissemination, supporting the view of Graves and Wells in their proposal of "factual accountability" (Graves and Wells, 2019). Instead of relying on specific entities for preventing the spread and creation of falsehoods, our work emphasizes that all actors involved have their roles in this fight—a form of joint accountability:

1. Users should become aware of their role in creating, disseminating, debunking, and preventing false information from being shared online. Online interventions could play a crucial role in circumventing any psychological factors that influence users' online behaviors (Pennycook et al., 2021).

2. Social media platforms should be able to take the role of fighting misinformation through both algorithmic and manual methods. Platforms could implement rapidly accessible interventions to prevent users from sharing misinformation (Epstein et al., 2021).

3. News media outlets, regardless of their political orientation, should comprehend their role in spreading both accurate and false information and promoting the former.

4. Political actors and organizations should invest in regulation to combat misinformation at its core through international collaborations.

5.2 Limitations and Future Work

Even though we have differentiated responsibility attribution across multiple aspects of online misinformation, we did not obtain respondents' attribution of specific roles and their relationships between entities. For instance, respondents might hold governments responsible for preventing misinformation through social media platforms' regulation, but not through restrictions to users' freedom of speech. Future work should delve deeper into these questions via structured interviews where research can address these questions in depth.

How and whether the public opinion should be embedded in future interventions and strategies to combat misinformation is an open question. Following the public opinion on this topic might prove to be difficult or even unproductive. Future discussions should consider descriptive research, such as ours, while weighting feasibility, practicality, and many other factors. Our studies' samples are also restricted to the US. Future research should be expanded to different communities, many of which have already suffered from misinformation (e.g., South America (New York Times, 2018) and Asia (Cha et al., 2020)).

Acknowledgments

G. Lima and M. Cha were supported by the Institute for Basic Science (IBS-R029-C2) and the Basic Science Research Program through the National Research Foundation of Korea (NRF-2017R1E1A1A01076400). We are grateful to several colleagues for their valuable feedback to this research.

References

Hon Sophia S Balod and Michael Hameleers. 2019. Fighting for truth? the role perceptions of filipino journalists in an era of mis-and disinformation. *Journalism*, page 1464884919865109.

Michael Barthel, Amy Mitchell, and Jesse Holcomb. 2016. Many americans believe fake news is sowing confusion. *Pew Research Center*.

BBC. 2019a. European elections: How disinformation spread in facebook groups. https://tinyurl.com/y2r8y2bm. Accessed 07 January 2021.

BBC. 2019b. 'i was a macedonian fake news writer'. https://tinyurl.com/y68lhmpy. Accessed 13 January 2021.

Joshua A Braun and Jessica L Eklund. 2019. Fake news, real money: Ad tech platforms, profit-driven hoaxes, and the business of journalism. *Digital Journalism*, 7(1):1–21.

Meeyoung Cha, Wei Gao, and Cheng-Te Li. 2020. Detecting fake news in social media: An asia-pacific perspective. *Communications of the ACM*, 63(4).

CNN. 2020. *How Twitter, Facebook and YouTube are handling election misinformation*. https://tinyurl.com/ycrztxfv. Accessed 04 January 2021.

Joan Donovan and Brian Friedberg. 2019. Source hacking: Media manipulation in practice. *Data & Society*.

Ziv Epstein, Adam J Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G Rand. 2021. Developing an accuracy-prompt toolkit to reduce covid-19 misinformation online. *Harvard Kennedy School Misinformation Review*.

Daniel Funke, Susan Benkelman, and Cristina Tardáguila. 2019. *Factually: How misinformation makes money*. https://tinyurl.com/yx8ktszt. Accessed 07 January 2021.

Lucas Graves and Chris Wells. 2019. From information availability to factual accountability: Reconsidering how truth matters for politicians, publics, and the news media. In *Journalism and Truth in an Age of Social Media*, pages 39–57. Oxford University Press.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent Features of Rumor Propagation in Online Social Media. In *Proceedings of the IEEE International Conference on Data Mining*.

Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *Data & Society*.

Melinda McClure Haughey, Meena Devii Muralikumar, Cameron A Wood, and Kate Starbird. 2020. On the misinformation beat: Understanding the work of investigative journalists reporting on problematic information online. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–22.

Amy Mitchell, Jeffrey Gottfried, Galen Stocking, Mason Walker, and Sophia Fedeli. 2019. Many americans say made-up news is a critical problem that needs to be fixed. *Pew Research Center*.

New York Times. 2018. Fake news is poisoning brazilian politics. whatsapp can stop it. https://tinyurl.com/y8adkq59. Accessed 07 January 2021.

Stefan Palan and Christian Schitter. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

The Atlantic. 2017. *What Facebook Did to American Democracy*. https://tinyurl.com/y7qgsgzw. Accessed 04 January 2021.

The New York Times. 2020. *Facebook and Twitter Dodge a 2016 Repeat, and Ignite a 2020 Firestorm*. https://tinyurl.com/y2w3wpgq. Accessed 04 January 2021.

The New York Times. 2021. *Russian Interference in 2020 Included Influencing Trump Associates, Report Says*. https://tinyurl.com/sv3jm94w. Accessed 27 May 2021.

The Walt Street Journal. 2016. *Mark Zuckerberg Continues to Defend Facebook Against Criticism It May Have Swayed Election*. https://tinyurl.com/ycz9thuh. Accessed 04 January 2021.

The Washington Post. 2019. Misinformation about covid vaccines is already spreading. it'll only get worse. https://tinyurl.com/y33jnms2. Accessed 07 January 2021.

The Washington Post. 2020a. *Twitter and Facebook warning labels aren't enough to save democracy*. https://tinyurl.com/ya6co7ub. Accessed 04 January 2021.

The Washington Post. 2020b. *Two things Facebook still needs to do to reduce the spread of misinformation*. https://tinyurl.com/y6hma674. Accessed 04 January 2021.

Yariv Tsfati, HG Boomgaarden, J Strömbäck, R Vliegenthart, A Damstra, and E Lindgren. 2020. Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. *Annals of the International Communication Association*, pages 1–17.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

E-BART: Jointly Predicting and Explaining Truthfulness

Erik Brand
The University of Queensland
e.brand@uq.net.au

Kevin Roitero
University of Udine
roitero.kevin@spes.uniud.it

Michael Soprano
University of Udine
michael.soprano@uniud.it

Gianluca Demartini
The University of Queensland
demartini@acm.org

Abstract

Automated fact-checking (AFC) systems exist to combat disinformation, however their complexity makes them opaque to the end user, making it difficult to foster trust. In this paper, we introduce the E-BART model with the hope of making progress on this front. E-BART is able to provide a veracity prediction for a claim, and jointly generate a human-readable explanation for this decision. We show that E-BART is competitive with the state-of-the-art on the e-FEVER and e-SNLI tasks. In addition, we validate the joint-prediction architecture by showing 1) that generating explanations does not significantly impede the model from performing well in its main task of veracity prediction, and 2) that predicted veracity and explanations are more internally coherent when generated jointly than separately. Finally, we also conduct human evaluations on the impact of generated explanations and observe that explanations increase human ability to spot misinformation and make people more skeptical about claims.

1 Introduction

Automated fact-checking (AFC) makes use of natural language processing (NLP) techniques to determine the veracity of a claim. The problem is defined in the following way: given a statement (claim) and some evidence, determine whether the statement is true with respect to the evidence (Stammbach and Ash, 2020). This is a challenging task for a human, let alone an autonomous system (Graves, 2018). However, AFC systems are able to approximate this process of evidence retrieval and synthesis with some degree of success (Stammbach and Ash, 2020; Vlachos and Riedel, 2014). The benefits and applications of an AFC system are numerous. The problem of disinformation is not new, however the rate of which it propagates has continued to increase, largely aided by the increasing popularity of social media platforms (Pennycook et al., 2021). AFC systems are starting to become a critical tool in combating the sheer quantity of claims that need to be verified.

While accurate (Stammbach and Ash, 2020; Portelli et al., 2020), AFC systems have been unable to supplement traditional fact-checkers due to a limitation in their design. A user may not accept to believe in a statement without first understanding the concepts and facts underpinning that statement. Such justifications are expected when reading journalistic fact-checking outcomes such as on Politifact; the fact-check outcome is accompanied by an explanation informing the reader of how the decision was reached. Without providing users with an explanation, the decision provided by an automated system is far less likely to be trusted (Toreini et al., 2020), especially as it is not generated by humans.

Automated systems have recently been developed to this effect, and have demonstrated promising initial results (Graves, 2018). While these initial results are unquestionably impressive, critical evaluation of the work reveals that many of these systems use separate models for veracity prediction and explanation generation. We argue that systems such as these are not actually describing their own actions and decision processes, and that the veracity prediction model is not made any more transparent.

In this paper, we propose and experimentally evaluate a system that jointly makes a veracity prediction and provides an explanation within the same model. This is novel as compared to classic post-hoc explainability methods that are built on top of existing machine learning models. As such, the generated explanations more closely reflect the decisions made by the veracity prediction model. In addition to this, we show that large transformer models are flexible enough to multitask, and are

Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 18–27,
Virtual, October 16-17, 2020.

thus able to explain their actions without detriment to the original task. This allows human end users to better interface with transformer models, fostering a more trustworthy relationship between humans and deep learning models.

We specifically address the following research questions:

- RQ1: How can we design a deep learning model to classify information truthfulness and, at the same time, generate a natural language explanation supporting its classification decision?
- RQ2: Can such model result in both accurate classification decisions and high quality natural language explanations?
- RQ3: Are machine-generated explanations useful for humans to better assess information truthfulness?

By creating an automated system that is capable of both evaluating the truthfulness of a statement and simultaneously generating a human-interpretable explanation for this decision, it is hoped that automated fact-checking systems will become more widely adopted.

2 Related Work

2.1 Existing Explainable-AFC Models

A number of techniques for generating explanations to accompany AFC decisions have been proposed. Saliency-based methods, such as those proposed by Shu et al. (2019) and Wu et al. (2020), use attention mechanisms to highlight the input that is most useful in determining the veracity prediction and present this information to the end user as a form of explanation. Logic-based approaches make use of graphs (Denaux and Gomez-Perez, 2020), rule mining, and probabilistic answer set programming (Ahmadi et al., 2019) to output a series of logical rules that result in a veracity prediction. This set of rules constitutes an explanation. While these methods are highly transparent and logical, the resulting explanation is not always human-readable (Ahmadi et al., 2019).

Summarisation techniques provide an explanation by summarising the retrieved evidence. The system proposed by Atanasova et al. (2020) utilises DistilBERT (Sanh et al., 2019) to pass contextual representations of the claim and evidence to two task-specific feed-forward networks which produce a classification and an extractive summary. Kotonya and Toni (2020) take a similar approach but tailor their model to the public health domain. The pipeline utilises Sentence-BERT (Reimers and Gurevych, 2019) to filter the evidence, a BERT-based veracity predictor, and a separate BERT-based summarisation model. The work by Kotonya and Toni (2020) differs from Atanasova et al. (2020) as it produces *abstractive* explanations, which are generally more coherent and similar to the way a human would generate a summary, rather than *extractive* explanations which take sentences verbatim from the evidence.

The framework proposed by Stammbach and Ash (2020) also produces abstractive explanations, but places higher emphasis on the evidence retrieval process. The framework consists of two components: 1) an evidence retrieval and veracity prediction module, and 2) an explanation generation module. The first component is an enhanced version of the DOMLIN system (Stammbach and Neumann, 2019), which uses separate BERT-based models for evidence retrieval and veracity prediction. For explanation generation, GPT-3 (Brown et al., 2020), a large pertained multi-purpose NLP model based on the Transformer, is used in 'few-shots' mode to generate a summary of the evidence with respect to the claim.

The system we present in this paper differs to the existing literature as rather than using two separate models for the veracity prediction and explanation generation, a single model is used to output both a veracity prediction and an abstractive summarisation.

2.2 BART Transformer Architecture

BART (Lewis et al., 2020) is a transformer (Vaswani et al., 2017) model that aims to generalise the capabilities of both BERT (Devlin et al., 2019) and GPT-style models. It consists of a bi-directional encoder, similar to BERT, as well as an auto-regressive decoder, similar to GPT. BART is pre-trained on a de-noising task whereby input text is corrupted and the model aims to reconstruct the original document, minimising the reconstruction loss. In contrast to existing de-noising models, BART is more flexible in that it is not trained to rectify a specific type of input corruption, but rather any arbitrarily corrupted document.

The pre-trained BART model can be fine-tuned to a number of downstream tasks. The authors

noted that the model performs comparably to other models, such as RoBERTa (Liu et al., 2019b), on natural language inference tasks. They also note that BART outperforms current state-of-the-art models on natural language generation tasks, such as summarisation (Lewis et al., 2020; Shleifer and Rush, 2020). Its ability to perform well on these two contrasting tasks made it an attractive choice as the base model for a system that can jointly predict the veracity of a claim, an inference task, and provide an explanation, a generative task.

3 A Model for Jointly Predicting and Explaining Truthfulness

Many of the systems in the reviewed literature use separate Transformer models for veracity prediction and explanation generation. Outlined here is our proposed architecture, E-BART, that jointly outputs a veracity prediction, as well as a human-readable, abstractive explanation addressing *RQ1*.

To adapt the BART-large encoder-decoder model to this downstream task, a 'joint prediction' head was developed. This head sits atop the BART model, and manipulates the transformer hidden states into the form of the desired output. Both the BART base model and the joint prediction head can be fine-tuned as a single unit to customise pre-trained BART weights to the joint prediction task.

The joint prediction head is depicted in green in Figure 1. The head takes as input the final decoder hidden state embeddings. It then passes all embeddings to a single feed-forward layer to produce a series of logits which form the basis of the predicted explanation. To facilitate classification, the hidden state embeddings corresponding to the final sequence separator token (`</s>` in BART) are extracted and passed to a small feed-forward network to shape the output to the desired number of classes. The logits obtained from this are then passed to a final soft-max layer to produce probabilities for each class. Unlike in BERT which uses embeddings corresponding to the `[cls]` token which is pre-pended to the input to perform classification, in BART the final sequence separator token is used instead as the decoder can only attend to the left of the current token. This conditions the classification on the entire input sequence. It is instructive to consider the training and inference processes separately, as they differ slightly due to the auto-regressive nature of the BART decoder.

During training, the encoder generates hidden

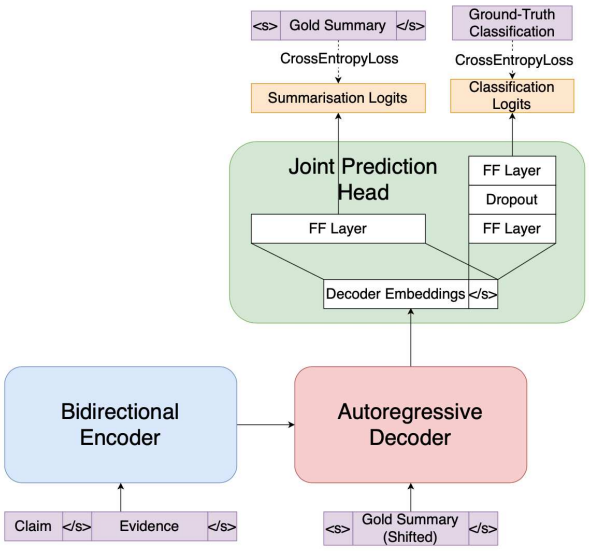


Figure 1: E-BART Training configuration.

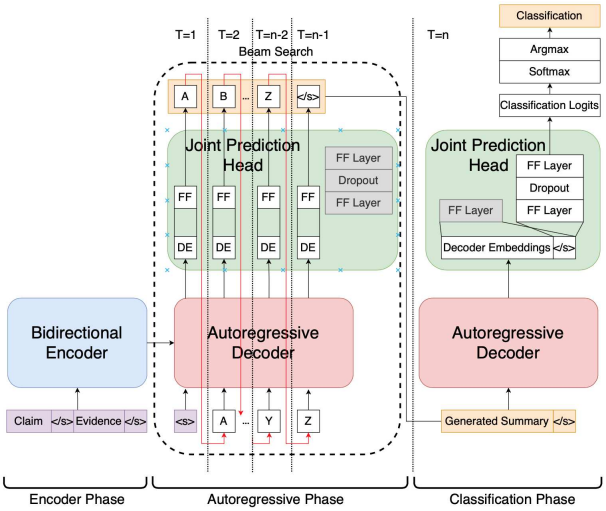


Figure 2: E-BART Inference process.

states from the tokenised input that are then injected into the decoder. The tokenised gold summary is presented to both the input and summarization output of the decoder, with the input shifted right by one token. This conditions the decoder to predict the next token given the current token. Concurrently, the classification labels are presented to the classification output of the joint prediction head. The loss is calculated as the weighted sum (with parameters α and $(1-\alpha)$) of the Cross Entropy Loss computed between the summarisation logits and the gold summary, and the Cross Entropy Loss between the classification logits and the ground truth classification.

Figure 2 shows the inference process. Running inference on the model begins by running

the encoder with the tokenised input to generate the encoder hidden states, as before. In contrast to the training process, the decoder is presented with the start sequence token (<s> in BART), and generates logits auto-regressively, guided by a beam search. The final phase of inference runs the decoder with the entire generated sequence presented at its input. At this point, the joint prediction head extracts the embeddings corresponding to the token immediately before the final sequence separator token from the generated sequence. This is done to mirror the training process. These embeddings are passed to the classification component of the joint prediction head, and then to a soft-max layer to produce the final classification.

4 Experimental Evaluation

4.1 Datasets

To evaluate the proposed models we make use of different datasets. The FEVER dataset consists of 185,445 claims, associated evidence, and veracity labels. The claims were generated by manipulating sentences taken from Wikipedia, and are labelled with either "Supports", "Refutes", or "Not_enough_info" based on whether the evidence entails the claim (Thorne et al., 2018).

The e-FEVER dataset by Stammbach and Ash (2020) augments the original FEVER dataset (Thorne et al., 2018) with explanations generated by their framework. It consists of 50,000 examples from the FEVER train set, and 17,687 from the development set. This provides a resource with claims, retrieved evidence, veracity labels, and explanations.

The e-SNLI dataset (Camburu et al., 2018) extends the SNLI dataset (Bowman et al., 2015) with human-generated explanations for each of the 570k examples. The SNLI task is to take two sentences and predict whether one entails, contradicts, or is neutral with respect to the other. e-SNLI adds complexity by also requiring a generated explanation for the label.

4.2 Training Methodology

To investigate *RQ2* and evaluate the performance of the proposed model on the FEVER and extended e-FEVER tasks, two different versions of the model were trained. In the e-FEVER dataset, if the GPT-3 component decided that the retrieved evidence was insufficient, it would produce a default 'null' explanation. Our first model, **E-BARTSmall**, was trained on the subset of the e-FEVER training set that did not include null explanations. This resulted in 40,702 examples. To process the data, the "+" character used to separate page titles from evidence was removed. The model inputs were tokenised and formatted as: "<s> claim </s> evidence </s>". The veracity labels were made numerical and explanations were tokenised in a similar manner. The processed dataset was used to fine-tune the BART-large model with joint prediction head for 3 epochs. Our second model, **E-BARTFull**, was trained in exactly the same way as the first, however it was trained using the entire e-FEVER training set, including examples with null explanations.

4.3 Evaluation Methodology

The development split of the e-FEVER dataset was prepared identically to the training split, producing e-FEVER_Full and e-FEVER_Small which do, and do not, include examples with null explanations, respectively.

When evaluating the veracity prediction accuracy of the models, it was noted that including the "Not_enough_info" class could under-represent the actual classification performance. Take the example in Table 1, which has a ground truth label of "Not_enough_info". Manual inspection shows that the explanation and evidence indicate that the claim is indeed refuted, which was correctly predicted by our model. Hence we report two sets of results, one with, and one without examples that have a e-FEVER label of "Not_enough_info".

4.4 Evaluation Results on Original FEVER

To compare with existing models, we report the classification performance of E-BART on the original FEVER development set. The DOMLIN system (Stammbach and Neumann, 2019) was used for evidence retrieval (discarding its veracity predictions) to provide evidence for 17k out of the 20k examples in the development set. We use our E-BART models to generate veracity predictions for the 17k examples, and then label the remaining with 'Not_enough_info,' as specified in the DOMLIN paper. Results are reported for the development set rather than the test set, as ground-truth labels were not published for the latter.

On the FEVER dataset, E-BARTSmall and E-BARTFull achieved label accuracies of **75.0** and **75.1**, respectively, outperforming state-of-the-art methods. For comparison, other published model

Table 1: Ground truth label is "Not_enough_info" and predicted label is "Refutes".

Claim	Evidence	Generated Explanation
Marnie was directed by someone who was "The Master of Nothing".	Alfred Hitchcock Sir Alfred Joseph Hitchcock (13 August 1899-29 April 1980) was an English film director and producer, at times referred to as "The Master of Suspense". Marnie (film) Marnie is a 1964 American psychological thriller film directed by Alfred Hitchcock.	Marnie was directed by Alfred Hitchcock, who was "The Master of Suspense".

accuracies on this dataset include: BERT-BASED 74.6 (Soleimani et al., 2020), DOMLIN 72.1 (Stammbach and Neumann, 2019), UCL MR 69.7 (Yoneda et al., 2018), UNC 69.6 (Nie et al., 2019), and UKP-Athene 68.5 (Hanselowski et al., 2018). E-BART compares favourably to the existing literature despite the e-FEVER training set having 95k less examples compared to FEVER, which the other models were trained on. It is hypothesised that the performance improvements are derived from using BART as a base model, and from requiring the model to further attend to the most relevant evidence in forming an explanation. The most noteworthy comparison is between E-BART and DOMLIN, which use identical evidence retrieval mechanisms, thus isolating the contribution of E-BART over standard veracity predictors.

4.5 Evaluation Results on e-FEVER

Table 2 shows the results obtained on the development e-FEVER dataset. To the best of our knowledge, there have been no other results reported on this recent dataset, hence we present a comprehensive snapshot of E-BART's performance.

Perhaps unsurprisingly, both our models performed better on e-FEVER_Small, which contained less inconclusive examples. More surprising is the consistency of E-BART's performance regardless of whether it was trained on e-FEVER_Small or e-FEVER_Full. This indicates that E-BART is robust to situations where evidence is sparse. Table 3, qualitatively shows that the model can even express the fact that it was not able to find relevant evidence.

The ROUGE metrics evaluate the consistency between the generated and e-FEVER dataset explanations, but are not necessarily representative of explanation quality. For instance, the explanation generated by GPT-3 may include some additional information compared to E-BART. Whether this additional information results in a better explanation compared to something more succinct is largely subjective and dependent on the system's use case. In Tables 1, 3 and 4, we present examples from the development set.

4.6 Evaluation Results on e-SNLI

The e-SNLI task presents a similar challenge to e-FEVER, whereby the entailment between two sentences is predicted (similar to predicting veracity of a claim with respect to evidence), and an explanation is generated.

A different version of the E-BART model was trained specifically on this dataset. The data was prepared by enumerating the labels, removing noisy data, and tokenising the summaries. The first and second sentences were concatenated and tokenised in the same way as the claim and explanation for the e-FEVER evaluation.

On the test e-SNLI dataset, E-BART achieved a label accuracy of **90.1** and a BLEU score of **32.70**. The model proposed in conjunction with the e-SNLI dataset, e-INFERSENT, achieved an accuracy of 84.0 and BLEU score of 22.4 (Camburu et al., 2018). In calculating the BLEU metric for the explainable models, the first two gold explanations were used as references as per (Camburu et al., 2018). As a further comparison, the following are the best performing models published in the literature[1] which, however, do not provide explanations: CA-MTL 92.1 (Pilault et al., 2021), SemBERT 91.9 (Zhang et al., 2020), MT-DNN 91.6 (Liu et al., 2019a), SJRC 91.3 (Zhang et al., 2018), D-CRCo-AN 90.1 (Kim et al., 2019), and LMTransformer 89.9 (Radford et al., 2018).

The performance of E-BART compares favourably with the state-of-the-art for this different task, corroborating the result from the FEVER task, and further confirming that E-BART

[1] https://nlp.stanford.edu/projects/snli/

Table 2: Effectiveness of the models on the e-FEVER dataset.

Model	Dataset	Accuracy no N.E.I	Accuracy full	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Sum
E-BARTSmall	eFEVER_Small	87.2	78.2	73.581	64.365	71.434	71.585
E-BARTSmall	eFEVER_Full	85.4	77.1	59.447	50.177	57.697	57.782
E-BARTFull	eFEVER_Small	87.1	78.1	64.530	55.283	62.691	62.820
E-BARTFull	eFEVER_Full	85.2	77.2	65.511	57.598	64.071	64.144

Table 3: Ground truth label is "Supports" and predicted label is "Supports".

Claim	Evidence	Generated Explanation
CBS is the network that aired The Millers	The Millers The multi-camera series aired from October 3, 2013 to July 18, 2015 and ran 34 episodes over two seasons on CBS. CBS announced the cancellation of The Millers on November 14, 2014, four episodes into the show's second season.	The Millers aired on CBS, however it does not say whether it was the network that aired it.

is flexible enough to provide explanations without compromising its performance on the main task. To compliment the BLEU metric, we provide some examples in Tables 5 and 6 for manual verification of explanation quality.

4.7 Validating the Use of Joint Models: Experiment 1

To further investigate *RQ2* and test the ability of our joint models, we compare the performance of E-BART to a pipeline that produces a veracity prediction and generates an explanation using two independent models. To ensure that the results between the two methods are comparable, the architecture and training methodology was kept as consistent as possible. The separate pipeline, which we term Separate-BART, uses a BART-based sequence classifier, and a BART-based model for language generation. Both E-BART and Separate-BART were initialised with the same pre-trained weights, and were trained and evaluated on e-FEVER_Small. However due to memory constraints, the inputs were truncated to a maximum length of 256 tokens (which only truncated 4.56% of examples). In addition to this, a virtual batch size of 32 was used (batch size four, with eight gradient accumulation steps) to overcome convergence issues. When training the sequence generator model, a batch size of two with two gradient accumulation steps was used, also due to memory restrictions on available hardware. In comparison, the joint model was trained with a batch size of four and no additional gradient accumulation.

The results in Table 7 indicate that the prediction performance of both types of model is almost identical, with Separate-BART being slightly more effective. Manual inspection of the generated explanations revealed that both were of a similar quality in terms of expressiveness and cohesiveness. This experimental result reinforces what was seen in the practical evaluations on e-FEVER and e-SNLI: that E-BART is able to jointly provide an explanation without diminishing the performance on its main task.

4.8 Validating the Use of Joint Models: Experiment 2

This experiment aims to investigate whether the internal consistency between the predicted veracity and predicted explanation differs between the joint and separate models. We use the same E-BART and Separate-BART models from Experiment 1, but train an additional 'judge' model to predict the veracity of a claim, given an *explanation*. The ground truth veracity labels and dataset explanations from e-FEVER_Small were used to train the BART-based sequence classifier. As such, its weights are not conditioned on those of E-BART or Separate-BART, meaning that it is independent from both models.

We run the experiment by taking the claims from the development set and the predicted explanations from E-BART. The claims and explanations are then passed to the 'judge' model to

Table 4: Ground truth label is "Refutes" and predicted label is "Refutes" (E-BART˙Small Model).

Claim	Evidence	Gold Explanation	Generated Explanation
Janet Leigh was incapable of writing.	Janet Leigh Janet Leigh (born Jeanette Helen Morrison; July 6, 1927-October 3, 2004) was an American actress, singer, dancer and author. Janet Leigh She also wrote four books between 1984 and 2002, including two novels.	Janet Leigh was capable of writing.	Janet Leigh wrote four books between 1984 and 2002, including two novels.

Table 5: Ground truth label is "Entailment" and predicted label is "Entailment".

Sentence 1	Sentence 2	Generated Explanation
One tan girl with a wool hat is running and leaning over an object, while another person in a wool hat is sitting on the ground.	A tan girl runs leans over an object	One tan girl with a wool hat is running and leaning over an object is a rephrasing of a tan girl runs leans over an object.

produce a veracity prediction. This 'judge' veracity prediction is then compared against the veracity prediction from E-BART, and the accuracy is computed. The process was repeated for Separate-BART, and the results are presented in Table 8.

The results show a higher accuracy for E-BART as determined by the 'judge' model. This provides indication that the veracity prediction and explanation generated by E-BART are more consistent with each other than those generated by Separate-BART. Ultimately this means that joint models are one step closer to being truly interpretable compared to models that generate explanations separately in a post-hoc manner. While this is not conclusive proof, it does provide some evidence that there are consistency gains to be made when using joint prediction and explanation models.

4.9 Testing the Impact of Explanations

To address *RQ3*, we experimentally validated the benefit of explanations generated by our model with human annotators, performing the crowdsourcing study detailed in the following. We collected the data using the Amazon MTurk crowdsourcing platform. To test the impact of machine-generated explanations of truthfulness, we deployed two versions of the same human annotation task. In the first version (Task 1), we provided participants with a claim from the FEVER dataset and we asked them to provide a truthfulness assessment on the true/false binary scale along with a sentence justifying their assessment, as this has been shown to improve assessment quality (Kutlu et al., 2020). In the second version (Task 2), we provided participants with both the claim and the explanation generated by our E-BART system and ask for an assessment and justification. Thus, the only difference between Task 1 and 2 is the absence or presence of the E-BART explanation. Each worker has been asked to assess the truthfulness of four claims, two labelled in the ground truth as "Supports", and two labelled as "Refutes". Each claim has been assessed by ten distinct human participants. To avoid bias, we performed a randomisation process while generating the claim-participant assignments (i.e., in the MTurk HITs). For consistency, we kept the same assignments (i.e., same HITs) for the two versions of the task. Participants were only allowed to complete one version of the task. To ensure high quality of the collected data and to avoid adversarial behaviour, we required participants to spend at least 2 seconds on each task page.

Figure 3 shows the external agreement between the ground truth and the crowd when considering both the individual participant judgments and the judgements aggregated over the ten participants assessing the same claim using majority vote. As we can see from the figure, the accuracy of the participants in Task 2 is higher than the ones of Task 1 (0.70 for Task 1 and 0.74 for Task 2 for the individual; 0.83 for Task 1 and 0.90 for Task 2 for the aggregated judgments). The differences are statistically significant according to the Mann-Whitney

Table 6: Ground truth label is “Neutral” and predicted label is “Neutral”.

Sentence 1	Sentence 2	Generated Explanation
A young family enjoys feeling ocean waves lap at their feet.	A young man and woman take their child to the beach for the first time.	Just because a young family enjoys feeling ocean waves lap at their feet does not mean that they take their child to the beach for the first time.

Table 7: Effectiveness of the joint and separate models.

Model	Accuracy no N.E.I	Accuracy full	Rouge 1	Rouge 2	Rouge L	Rouge Sum
E-BART	87.2	78.2	73.581	64.365	71.434	71.585
Separate-BART	88.1	78.9	73.070	63.634	71.005	71.136

Table 8: Internal consistency of the joint and separate models.

Model	Accuracy no N.E.I	Accuracy full
E-BART	91.8	86.8
Separate-BART	90.4	85.8

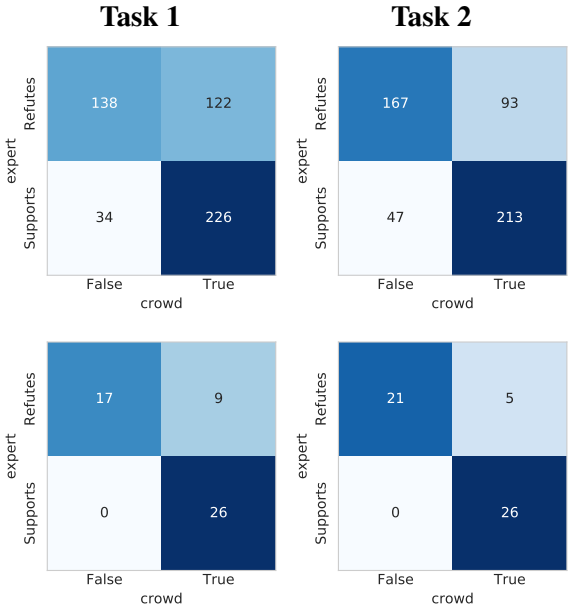


Figure 3: External agreement between ground truth and crowd for raw (first row) and aggregated (second row) truthfulness assessments. Task 1 shows just the claim while Task 2 shows the claim and the natural language explanation generated by our E-BART model.

U test at the $p < 0.05$ level for both the individual and the aggregated judgements. We can additionally observe that the display of explanations (i.e., Task 2) reduces the number of *false positives* (i.e., claims that are false but are erroneously perceived as being true by human subjects) from 122 to 93; Thus, it appears that the explanations automatically generated by our E-BART model have the effect of making people more skeptical about claims (see also Table 3 for an example). Performing simple aggregations and under condition of Task 2, we are able achieve 90% non-expert label accuracy, which is a promising step towards crowdsourced truthfulness annotations (Roitero et al., 2020).

5 Conclusions

In this paper we explored the potential of AFC models jointly making a prediction and providing a human-readable explanation for that prediction. To this end, we proposed the E-BART architecture and evaluated its performance on the extended FEVER and SNLI tasks. Experimentation revealed that E-BART could achieve results comparable to the state-of-the-art and simultaneously generate coherent and relevant explanations. We argued that jointly predicting explanations makes AFC systems more transparent, and fosters greater trust in the system. Finally, human evaluation of the impact of generated explanations revealed that the explanations provided by E-BART generally make people more accurate in detecting misinformation and more skeptical of a claim they encounter online.

Acknowledgments. This work is supported by a Facebook Research award, the ARC Discovery Project (Grant No. DP190102141), and by the ARC Training Centre for Information Resilience (Grant No. IC200100022).

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable Fact Checking with Probabilistic Answer Set Programming. In *Proceedings of the 2019 Truth and Trust Online Conference*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, pages 9560–9572, Montréal, Canada.

Ronald Denaux and Jose Manuel Gomez-Perez. 2020. Linked Credibility Reviews for Explainable Misinformation Detection. In *The Semantic Web – ISWC 2020*, pages 147–163. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, USA. Association for Computational Linguistics.

Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-Checking. *Reuters Institute for the Study of Journalism*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic Sentence Matching with Densely-Connected Recurrent and Co-Attentive Information. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 6586–6593, Honolulu, Hawaii, USA. AAAI Press.

Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7740–7754. Association for Computational Linguistics.

Mücahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator Rationales for Labeling Tasks in Crowdsourcing. *Journal of Artificial Intelligence Research*, 69:143–189.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866, Honolulu, Hawaii, USA.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Jonathan Pilault, Amine Elhattami, and Christopher J. Pal. 2021. Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data. In *Proceedings of the 9th International Conference on Learning Representations*.

Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus. 2020. Distilling the Evidence to Augment Fact Verification Models. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 47–51. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understandingby Generative Pre-Training.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3980–3990, Hong-Kong, China. Association for Computational Linguistics.

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR Conference*, page 439–448, New York, NY, USA. Association for Computing Machinery.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained Summarization Distillation. *arXiv preprint arXiv:2010.13002*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference*, page 395–405, New York, NY, USA. Association for Computing Machinery.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for Evidence Retrieval and Claim Verification. In *Proceedings of the 42nd European Conference on IR Research*, volume 12036, pages 359–366, Lisbon, Portugal. Springer.

Dominik Stammbach and Elliott Ash. 2020. e-FEVER: Explanations and Summaries for Automated Fact Checking. In *Proceedings of the 2020 Truth and Trust Online Conference (TTO 2020)*, page 32. Hacks Hackers.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)"*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 272–283, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA, USA.

Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035. Association for Computational Linguistics.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL Machine Reading Group: Four Factor Framework For Fact Finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Zuchao Li, and Hai Zhao. 2018. Explicit Contextual Semantics for Text Comprehension. *arXiv preprint arXiv:1809.02794*.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 9628–9635, New Your, NY, USA. AAAI Press.

The Emergence of Deepfakes and its Societal Implications: A Systematic Review

Dilrukshi Gamage
Department of Innovation Science,
Tokyo Institute of Technology,
Tokyo, Japan
dilrukshi.gamage@acm.org

Jiayu Chen
Department of Psychology and Human Developmental Sciences,
Nagoya University,
Nagoya, Japan
chen.jiayu@h.mbox.nagoya-u.ac.jp

Kazutoshi Sasahara
Department of Innovation Science,
Tokyo Institute of Technology,
Tokyo, Japan
sasahara.k.aa@m.titech.ac.jp

Abstract

The appearance of Deepfake tools and technologies in the public is proliferating. Scholarly research is very centered on technology of deepfake but sparse in understanding how the emergence of deepfakes impacts society. In this systematic review, we explored deepfake scholarly works that discuss societal implications than the technology-centered focus. We extracted studies from major publication databases - Scopus, Web of Science, IEEEXplore, ACM Digital Library, Springer Digital Library and Google Scholar. The corpus reflects patterns based on their research methodologies, area of focus, and the distribution of such research. Out of 787 works, 88 were highly relevant, with the majority of the studies being reviews of the literature. While research focus is generally drawn upon exploring security related harms, less focus is put on issues such as ethical implications and legal regularities for areas other than pornography, psychological safety, cybercrimes, terrorism, and more. The field research for Deepfake social impact research is emerging and this paper brings more insights drawn from a methodical, subject focused and distribution point of view.

1 Introduction

The rapid development of technologies such as Artificial Intelligence (AI) and Deep Learning (DL) revolutionized the way we create and consume content. As a byproduct of this revolution, we witness emerging technologies such as Deepfake which may potentially harm and distress social systems. Deepfakes are synthetic media generated using sophisticated algorithms which reflect things that did not happen for real but computer generated for manipulation purposes (Westerlund, 2019). In many cases, specific methods of Deep Learning which involve training generative neural networks — autoencoders, Generative Neural Network (GNN) in Machine Learning (ML) are utilized to generate these synthetic media.

Currently, a myriad of scholarly works concentrate on specific Deep Learning techniques — types of neural network model in which the model is trained to restore (copy) the input data known as auto encoders, GAN Models that involves a generator and discriminator in building an image closer to the original, High-definition face image generations, Conditional GANs (CGAN) that generate data while controlling attributes by giving attribute information in addition to images during training, face swapping techniques and speech synthesizing techniques (Guarnera et al., 2020). These studies are more influenced by the Deepfake generation and detection methods. However, the advancements of these scholarly works and the democratization of these technologies made it easy for any individual to generate realistic fake media content which could have been difficult previously. Apart from the incident that incepted Deepfake in 2017 where celebrity faces were used to create phonographic videos using Deepfake technologies (Burkell and Gosse, 2019), the incidences such the British energy company scammed by voice Deepfake technology (Stupp, 2019) in 2019 and recently the arrest of a Japanese student for posting pornographic videos that synthesized the face of a celebrity using Deepfake technology by training the model for about a week, using 30,000 images per video where the case is believed to be the first criminal case in Japan which Deepfake technology was abused (Times, 2020) can be highlight as emerged abuse of using Deepfakes. In addition to these, more recently(March 10th 2021), a mother in Pennsylvania used Deepfake technology to forge photos and videos to show drinking, smoking and nakedness to trap a teammate of a high school daughter who works as a cheerleader (Guardian, 2021) and the article written in

Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 28–39,
Virtual, October 16-17, 2020.

Newyorker inquires ethical implications of Deepfake voice by narrating the movie about celebrity chef Anthony Bourdain in July '15th, 2021 (Ronsner, 2021). Together, all such incidences have demonstrated the emerging threats unresting the social process.

Although Deep Learning technologies are versatile and could be useful in revolutionizing various industries, these incidents collectively raise concerns about the societal problems emerging from them. There is ample work in computer science on automatic generation (Yadav and Salmani, 2019; Caldelli et al., 2021) and detection of Deepfakes (Maksutov et al., 2020; Rana and Sung, 2020), but to date there are only a handful of social scientists who have examined the social impact of Deepfake technology. In this paper, we conducted a systematic literature review to understand the existing landscape of research that examines the possible effects Deepfakes might have on people, to understand the psychological dynamics of deepfakes and to discover how it impacts society. In particular, we hope to examine the following two research questions:

- Q1: What types of research conducted between 2017-2021 to understand the psychological and social dynamics and societal implications of Deepfake?
- Q2: What is the distribution of Deepfake research between 2017-2021 that explores any type of psychological dynamics and its societal implications?

The objective of this systematic study is to highlight the types of research carried out to understand the social dynamics of Deepfake and identify any gaps in the researches that need further discussions on social implications and concerns that arise from the technology. This exploration of research related to social processes and the implications of Deepfake will provide necessary projections, and point to scholarly work in this area where social scientists could make a useful contributions by understanding any lack of new directions. Since deepfake attributes in Deep Learning and Machine Learning, much advancement and research has occurred in the field of computer science. In addition, with the democratization of accessible technology to a wider audience, necessary attention is paramount in order to understand the societal implications of this phenomenon.

Search Database	Hits	Selected
Springer Online Database	177	17
IEEE	154	11
ACM	264	8
Web of Science	137	41
Scopus	55	2
Other (Google Scholar)	NA	9
Total	787	88

Table 1: Summary of the results retrieved by running the search query and manually filtering by reviewing according to the inclusion criteria.

2 Methods

We obtained articles for our systematic review by searching popular scientific search engines and repositories—Springer Digital Database,IEEEXplore, ACM Digital Database, Web of Science, and Scopus. Most systematic reviews incorporate Preferred Reporting Items for Systematic reviews and Meta-Analyses protocols (PRISMA) explained in details by Moher et al. (2015).We followed a similar structure to this literature review with particular interest in understanding the two previously mentioned research questions. We used the following search query in all 5 databases and in addition to this, used Google Scholar to search any other relevant preprints or non-peer reviewed articles to bring more inclusively to the research which may not have been listed in ACM, Scopus, IEEE, Web of Science or any other database.

{Deepfake OR Artificial Intelligence}
AND Misinformation

We did not restrict our search to only journal papers, but allowed any peer reviewed paper, or commentary in an article, critical review or even work-in-progress papers including the preprints. After the search terms provided the dataset, we used two experienced researchers to filter the research based on an inclusion criteria, we were particularly careful to select the results only if the manuscripts examined perceptions of Deepfake or its impact to human interaction or discussed the social implications of Deepfakes. In other words, articles that discussed a pure technology perspective (such as GAN), or studies to find new techniques for Deepfake detection's were eliminated as irrelevant to this study. Figure 1 describes the process conducted to obtain the relevant data to the analysis.

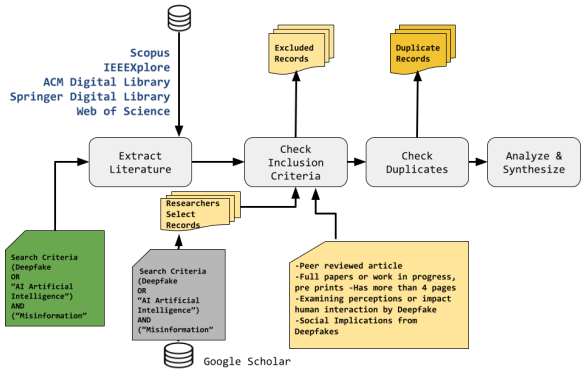


Figure 1: Flow of the systematic review

2.1 Dataset

Our initial search query extracted 787 articles from 5 databases. The extracted results were then combined to a single data file and two researchers collectively further filtered based on the inclusion criteria depicted in Figure 1 by manually reviewing the abstracts. In addition to these filtered articles, additional papers were added based on the relevant research found by Google Scholar and we labeled this source as "Other". Although a Google Scholar advance search returned 3420 hits, given the depth and spread of the articles we focused only on the first 20 pages which had 200 hits and selected 9 highly relevant papers not included in any databases.Out of these, 4 papers were from journals and, 2 universities repositories which was not listed in any of the 5 databases. Another 2 were preprints and currently under review, 1 commentary from Nature. We found 79 highly relevant papers from the 5 original databases and with the Google Scholar results had 88 papers selected for analysis.A breakdown is depicted in Table 1.

2.2 Measures

To answer RQ1, we analyzed all 88 papers using their full text, summarized the key phrases, highlighted major findings in the respective papers and identified any themes under which the article could be categorized. Based on the summary and key phrases, it was evident that the corpus can be categorized by a common methodological standpoint. For example, we realized that each article can be categorized by whether it conducted an experiment to understand social dynamics or had any sort of methodical analysis to understand social impact or if it was produced as a result of an extensive critical review by positioning any premises or even if it provided a conceptual proposal or framework beyond the review of the Deepfake social phenomenon. At the same time, we also examined whether or not the corpus focused on several domain areas addressing Deepfake social issues. We incorporated word clouds on each abstract to support subjective judgment on categories and focus areas.

To answer the RQ2, on the distribution of research in Deepfake psychological dynamics and its societal implications, we described descriptive statistics with a network analysis that understands the connections with its type of research and emphasis. At the same time, to highlight the emphasis of the paper, we highlighted the generated word clouds, specifically depict the categorical flows based on the frequencies, and used the network diagram using Gephi software to illustrate the author distributions among the selected papers.

3 Results and Discussion

Overall,the majority of the results from the query resulted scholarly work related to Deep learning, AI and ML learning technologies, and its improvements in creating or detecting Deepfake. Only 88 out of 787 were selected as those research works were found to be discussing the psychological dynamics, social implications, harms to the society, ethical standpoint, and or solutions from the a social-technological point of view.

3.1 RQ1: Types of research

Examining the abstracts and full text of the articles, we identified that each article could be categorized based on 11 types of research—Systematic review, Review based on Literature, Philosophical mode of enquiry, Examines, Experiment, Network Analysis, Content Analysis, Design, Conceptual Proposal, Commentary and Analysis by Examples. Although these categories are based on the subjective judgment of the authors, it provides a solid understanding to the conducted research based on its main objectives and methods.

A magnified view of this dataset (88) revealed that the majority (30) of the papers focused on critical reviews based on the previous literature and slightly above half of the papers (21) conducted active experiments using real users to explore the social and psychological dynamics of perceiving Deepfakes or understanding their impact. Only one study was performed a network analysis based

on Deepfake discourse and limited other research papers focused on rest of the methods as depicted in Figure 2. Apart from the methodology point of view, we also derived key categories of the papers based on its focus area. Although our key interest centered upon Deepfake and its social impact, we observed that these relevant research covered a wider range of focus areas in different subject domains. These areas ranged from security aspects, pornography, legal concerns, Deepfake media, specifically video and images, psychological perspectives, political perspective, human cognition perspectives, and more. Therefore, to specifically answer RQ1, we describe the details of these methodologies and focus areas in the following sections.

Methodology used in Deepfake social implication research

Although methodical approaches for research are not new, our analysis of the 88 highly relevant papers for the social or psychological implications of Deepfake reflected that most of the research in this domain is still developing and many researchers are critically evaluating and analyzing Deepfake phenomena from the previous literature, discussing potential future outcomes. We categorized this type of research as **Review based on Literature** and from our corpus, the earliest research on critical reviews of Deepfake social implications occurred in 2019 (although the term "Deepfake" first time in 2017 (Westerlund, 2019)). Research by Westling (2019) raise questions about to understanding whether the Deepfake phenomena is shallow or deep and how society might react to these technologies. Specifically the paper critically analysed and predominantly provided nuances to the technology that generates deep fake media and its uses, showing that society has never relied solely on the content as a source of truth.

Similarly Antinori (2019) provides an extensive narration to Deepfake and relates its consequences to terrorism. The author does not follow a systemic approach, however there is a critical discussion of the Deepfake focus on the near future of security threats by using examples of previous literature and emphasizing the need of awareness, law enforcement, and policymakers to implement effective counter terrorism's strategies. While providing this background and previous work, the author also articulates his stance on the subject emphasizing that as a globalized community, we are transitioning from e-terrorism to upcoming online terrorism, as well as the linearity to hypercomplexity by malicious use of AI and living in the post-truth era of a social system. Since his research article not only provides critical review based on past literature but also the authors theoretical and qualitative research experience with participation and working as a counter terrorism expert in related projects, we also intersected this with a new category: **Examines**. Through our full-text analysis, we observed that many other **Review based on Literature** scholarly work intersects with the **Examines** category. In these types of articles, we observed authors critically providing their experience or using their point of view as a metaphor to build constructs. All together we found 11 out of 30 papers categorized as Review based on Literature illustrated this intersection. For example, the review article by Hancock and Bailenson (2021), attempts to understand the possible effects Deepfakes might have on people, and how psychological and media theories apply. In addition, the article by Öhman (2019) brings a philosophical mode of enquiry to a pervert's dilemma, an abstraction about fantasizing sexual pornography and argues that ethical perspectives underline dilemmas by using the literature and theories. Similar placement of arguments and concepts supported by review of literature can be found in articles by Taylor (2021), Kerner and Risse (2021), Langa (2021), Ratner (2021), Harper et al. (2021), Langguth et al. (2021) and (Greenstein, 2021). However, we also derived 4 research articles that falls in the category of **Examines** without a dominating critical literature review—For example, an article was compiled while examining US and British legislation indicating legislative gaps and inefficiency in the existing legal solutions and presenting a range of proposals of legislative change to the constitutional gaps in porn (Mania, 2020). The article examines current online propaganda tools in the context of the different information environment and, provides examples of its use, while seeking to educate about Deepfake tools and the future of propaganda (Pavlíková et al., 2021). Another study examines the problem of unreliable information on the internet and its implications for the integrity of elections, and representative democracy, in the U.S. (Zachary, 2020) and another study that addresses the economic factors

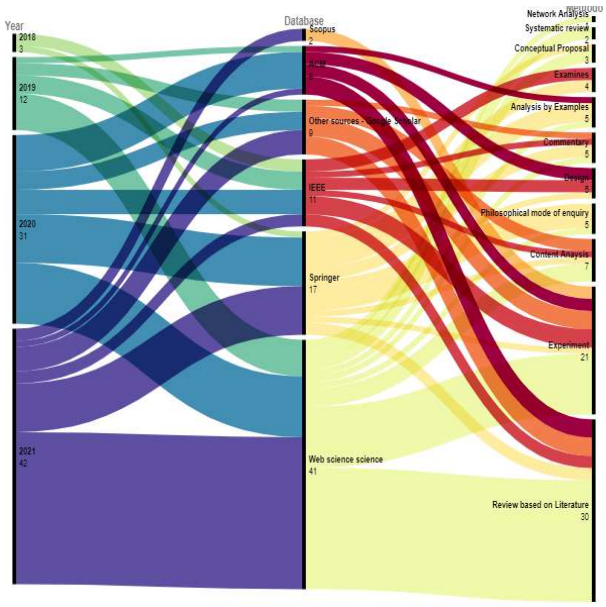


Figure 2: Scholarly work distribution based on the year it was published, the Published databases and its Methodology

that make confrontational conversation more or less likely in our era and brought viewpoints in the Deepfakes which becoming more widespread on the dark web (Greenstein, 2021) are falling in to this *Examines* category.

However, alongside review based articles and articles that conducted extensive examinination, we also derived another category. Although this category is similar to the methods we previosuly stated, it is distinguished by the way it positions its point of views. We noticed that this type of articles is extensively based on use cases, examples of incidences or more descriptions of theoretical and informational AI and Deepfake technologies. We name this category **Analysis by Example** and found 5 papers fall under its umbrella. Articles in this category includes Pantserev (2020), through their examples of Deepfakes in the modern world, and the internet-services, Amelin and Channov (2020) study the use of legal regulation in use of facial processing technologies, and Caldwell et al. (2020) study possible applications of artificial intelligence and related technologies in the perpetration of crimes, Degtereva et al. (2020) studied the general analysis of risks and hazards of the technologies and analysis examples of legal remedies available to victims. We also identified a category named **Philosophical Mode of Enquiry** which includes papers that use a philosophical point of view in premising their enquiry to the social issues found with in the Deepfake applications (Öhman, 2019; Ziegler, 2021; Floridi, 2018; Hazan, 2020; Kwok and Koh, 2021).

However, since the developments in the area of social implications of Deepfakes are yet growing, we observed only 2 **Systematic Review** types of research that explain in detail of the growing body of literature and its systematic analysis (Godulla et al., 2021; Westerlund, 2019). The first systematic review used English-language deepfake research to identify salient discussions; and the other used 84 publicly available online news articles to examine what deepfakes are and who produces them, and the benefits and threats of deepfake technology in 2021 and 2019 respectively. However, apart from these critical reviews, examiner papers, analysis by examples and systematic reviews, we found one other methods that could be classifed into the same theme but distinct in its narration of the information as it is made as a personal opinion or commentary to certain events. We named this category as **Commentary Bases** which often provides short narrative for the question of the future of technological implications (Kalpokas, 2021; LaGrandeur, 2021; Beridze and Butcher, 2019; Strickland, 2018, 2019).

As a next category of methodology, we observed that 21 out of 88 papers depicted some sort of experiment using human subjects to understand any impact and social implications of Deepfake and we named this category **Experiment**. In this category we observed researchers such as Khodabakhsh et al. (2019) used 30 users to examine human judgment on Deepfake videos, Caramancion (2021) used 161 users to explore the relationship between a person's demographic data, political ideology and the risk of him/her falling prey to Mis/Disinformation attacks. The largest study conducted by Yaqub et al. (2020) used 1,512 users to explore the impact of four types of credibility indicators on people's intent to share news headlines with their friends on social media. Similarly, Dobber et al. (2021) studied effects on political attitudes using 271 users, Köbis et al. (2021) studied the inability of people to reliably detect Deepfakes using 210 users. Their research particularly found neither by educating or introducing financial incentives improves their detection accuracy experimented and many other similar studies contained in this category. Apart from experiments, we also found research articles proposing frameworks or

solutions to Deepfake societal issues by conceptualizing theoretical frameworks (Cakir and Kasap, 2020; Kietzmann et al., 2020b,a) named as **Conceptual Proposals**. Beyond conceptual proposals, we also found that some articles consisted clear design goals with implementation plans or some some artifacts designed as solutions to the issues of Deepfake societal issues (Chi et al., 2020; Qayyum et al., 2019; Chen et al., 2018; Sohrawardi et al., 2019; Inie et al., 2020). Thus we introduced a category named **Design**.

Apart from such dominated methods to observe social implications and perceptions of Deepfakes, we also found 7 articles that followed the **Content Analysis** method. Three used Twitter data as their corpus (Maddocks, 2020; Oehmichen et al., 2019; Hinders and Kirn, 2020) and two studies analyzed the article content in news media (Brooks, 2021; Gosse and Burkell, 2020); each study conducted analyses using YouTube comment discourses about Deepfakes (Lee et al., 2021) and journalist discourse (Wahl-Jorgensen and Carlson, 2021) to understand the social implications of the Deepfakes phenomenon. Althogh, similar to these studies, we categorized one more study as **Network Analysis** and it conducted semantic content analysis using Twitter data relating to Deepfake phenomena (Dasilva et al., 2021) to understand the social discourse.

Range of focus areas examining Deepfake and its social implications

Apart from the key categorization towards research methods, we examined the significant research questions these research methods are used to solve. This aids us in categorizing the Deepfake social research based on the subject areas which it is focused. We derived 30 main focus areas these research articles primarily concentrate on, followed by 44 sub-focused areas. This flow is graphically represented in the alluvial diagram in Figure 3. At the interest of space for this paper, we highlight the top 5 focus areas of research.

As it appears, the highest interest of focus is drawn upon **Security** related issues relating to the social implications of Deepfakes. A significant number of research relating to security are foreseeing harms and threats to the society through "Review of literature" (Repez and Popescu, 2020; Taylor, 2021; Kaloudi and Li, 2020; Rickli and Ienca, 2021). More security focus research is conducted based on a "Design" of a blockchain-based

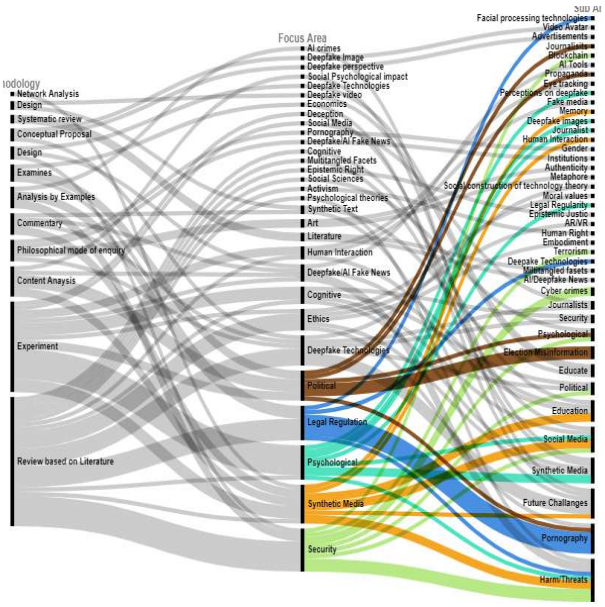


Figure 3: All 88 papers are categorised based on its main methodological and focus are of the research. Highlighted in color are the first five focus areas based in the higher frequency - Security, Synthetic Media, Psychology,Legal Regulation and Political are the top 5 focus areas.

framework for preventing fake news while introducing various design issues (Chi et al., 2020). At the same time security focus research has been visible in the research method of "Analysis by Example" where Degtereva et al. (2020) conduct a general analysis to understand the risks and hazards of the technologies used today and highlight the need for a wider application and enhancement of Deepfake technology to fight Cyber Crimes. Similarly, Pantserev (2020) analyses a wide range of examples of deepfakes in the modern world and the Internet-services that generate them with a key focus on security. Their research also depicts a clear sub focused area of **Psychological Security** as they try to understand the threats Deepfake cause to society and its impacts.

The next highest focus area of literature solves problems relating to "Synthetic Media." These are mostly considered as the Deepfake in the the mode of Videos. We observed that most researchers have used **Synthetic media** to conduct "Experiments" and "Content Analysis." For instant, Iacobucci et al. (2021) test whether a simple priming of deepfake information significantly increases users' ability to recognize Synthetic media, Hwang et al. (2021) examined the negative impact of deepfake video and the protective ef-

fect of media literacy education; and Murphy and Flynn (2021) examined how Deepfake videos may distort the memory for public events, yet found it may not always be more effective than simple misleading text. Other than these, Brooks (2021) used "Content Analysis" to analyze popular news and magazine to understand impact of Synthetic media. Interestingly, the article argues, that if fake videos are framed as a technical problem, solutions will likely involve new systems and tools or if fake videos are framed as a social, cultural, or as an ethical problem, solutions needed will be legal or behavioral ones. On the other hand, in this article, the focus of Synthetic media also expand to the sub focus to examine the societal **Harm/Threats**. Similarly, Hinders and Kirn (2020), empathize that digital photos are so easy to manipulate, yet deepfake videos are more important to understand as deepfake synthetic media (video evidence) could be deliberately misleading and not easy to recognize as fake. Apart from content analysis, focus on synthetic media narrowed the focus for a few commentary based articles: one examines Deepfake video implications on Facebook (Strickland, 2019), and two other articles focus examining Deepfake videos challenges with a sub focus on understanding **Future Challenges** (Kalpokas, 2021; LaGrandeur, 2021).

The next highest set of research articles focus mainly on the areas of **Psychological, Legal Regulation, and Politics**. Interestingly, all **Psychological** focus research conducted as experiments except for one that focuses on the Psychological impact of Deepfake through a review of literature (Hancock and Bailenson, 2021). In experiments, Yaqub et al. (2020) explore the effect of credibility signals and how they perceived any individual to share fake news Khodabakhsh et al. (2019) focus on understanding the vulnerability of Human judgement to Deepfake. Ahmed (2021b) examines the social impact of Deepfakes using an online survey sample in the United States. This investigates psychological aspects of the impact of Deepfake while examining the concerns of citizens regarding deepfakes, exposure to deepfakes, inadvertent sharing of deepfakes, the cognitive ability of individuals, and social media news skepticism. Cochran and Napshin (2021) provided psychological aspects of Deepfakes by exploring factors impacting the perceived responsibility of online platforms to regulate deepfakes and provide implications for users of social media, social media platforms, technology developers, and broader society. The research focusing on **Legal Regulation** extensively worked on Deepfake pornography, discussing its ethical perspective, consequences, and legal framework to take action (i,e '(Karasavva and Noorbhai, 2021; Delfino, 2020; Gieseke, 2020). Few others had sub-focus on discussing the threats and harms (O'Donnell, 2021), Terrorism (Antinori, 2019) and specific to facial processing technologies (Amelin and Channov, 2020). The **Political** focus researches have been extensively worked on election related consequences of Deepfakes and few focused on the journalists discourse to shape political context (Wahl-Jorgensen and Carlson, 2021), explored the relationship between political and pornographic deep fakes (Maddocks, 2020) and discussed the threat of Deepfake online propaganda tools (Pavlíková et al., 2021).

Figure 4: Word clouds from abstracts identified as focusing Pornography (top) and in all articles (bottom)

3.2 RQ2: Distribution of the research

In the previous sections, we partially stated the distributions of research methods and focus areas by utilizing Figure 2 and 3. Further, we expanded the knowledge of the landscape for Deepfake research that concentrates on its societal impacts by examining the yearly distribution of the relevant research. As depicted in Figure 2, the yearly projection reflects a trend for studies which explore the social implications by Deepfake are emerg-

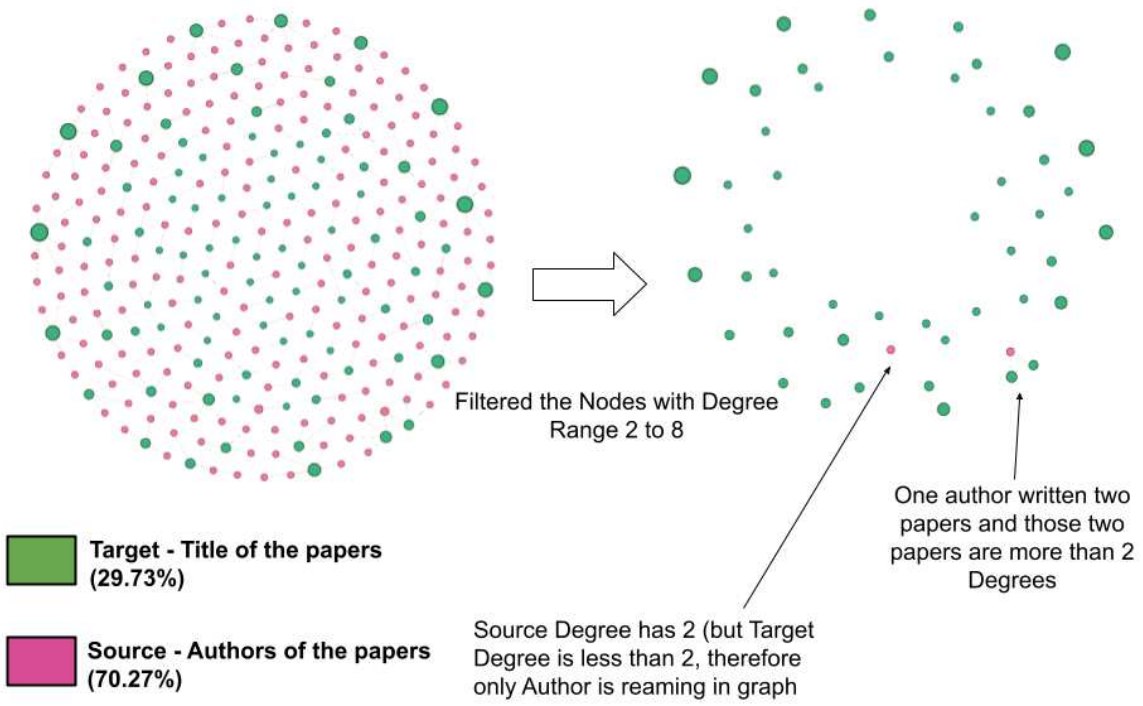


Figure 5: [Left] A bipartite graph created using source as the authors and targets as the papers. [Right] The Bipartite graph filtered based on the degree centrality larger than 2.

ing since 2019 and 2021 has the highest number of such researches(42) even before the year 2021 ends.

We generated word clouds for each abstract and one common word cloud combining all 88 abstracts to make sense of what we examined and to summarize the analysis of the full text of the articles. The top word cloud in the Figure 4 generated from a abstracts which we categorized as Pornography (Gieseke, 2020) and it hows its words are cantered on pornography; The bottom shows the word cloud from all abstracts which reflects Deepfake as the central theme and yet highlights, other focus areas we identified that greatly resonated in our categorizations. Finally, to better understand the distribution of the authors of these papers, we generated bipartite networks using the author list with the titles of the papers they have written (Figure 5). Nodes represent the authors (pink), papers (green), and the edges point from the authors to the papers. It appears that researchers who explore Deepfake social implications are almost not connected to each other as the clustering coefficient indicates 0.0 and nearly 30% of Papers written by 70% of authors and the highest number of relationship consisted one degree as a single author has written the papers. Ranked by the degree centrality (how many authors written how many papers), the graph revealed the lowest degree centrality as 1 and the highest as 8. Filtering the network to reflect if there are any 2 or more authors collaborated in writing these social research types we filtered the graph into 2 to 8 degree centrality. Interestingly, this resulted only two authors had 2 degrees relationship. in one instance, the same author wrote two different papers while collaborating with multiple other authors (Kietzmann et al., 2020b,b); in the other instance the same author has written two papers without any author collaborations (Ahmed, 2021c,a).

4 Conclusions

Our study reflects a comprehensive review of Deepfake research which discusses the social implications of Deepfake as the primary focus opposed to the reviews to the technology itself. We selected 88 highly relevant papers to our study and based on the methodical aspects, we found 11 types of studies that could be categorized. Out of all 88 papers, we also found that majority of studies focus on research relating to security and discuss the possible harms and threats to the social

echo system. Much debated issues such as ethical implications to Deepfake, the regulatory or legal solutions other than pornography, such as making awareness or educative activism to other type of harm specially, the cyber crimes and terrorism are much sparse in the landscape. Our results suggest that the social science of Deepfakes is emerging, but such research has been conducted independently thus far. Given that Deepfakes and related AI technologies are weaponizing, the social implications of Deepfakes should be more investigated with an interdisciplinary effort.

Acknowledgments

This work is generously supported by JST, CREST Grant Number JPMJCR20D3, Japan.

References

Saifuddin Ahmed. 2021a. Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences*, 182:111074.

Saifuddin Ahmed. 2021b. Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *new media & society*, page 14614448211019198.

Saifuddin Ahmed. 2021c. Who inadvertently shares deepfakes? analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, 57:101508.

Roman Amelin and Sergey Channov. 2020. On the legal issues of face processing technologies. In *International Conference on Digital Transformation and Global Society*, pages 223–236. Springer.

Arije Antinori. 2019. Terrorism and deepfake: From hybrid warfare to post-truth warfare in a hybrid world. In *ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics*, page 23. Academic Conferences and publishing limited.

Irakli Beridze and James Butcher. 2019. When seeing is no longer believing. *Nature Machine Intelligence*, 1(8):332–334.

Catherine Francis Brooks. 2021. Popular discourse around deepfakes and the interdisciplinary challenge of fake video distribution. *Cyberpsychology, Behavior, and Social Networking*, 24(3):159–163.

Jacquelyn Burkell and Chandell Gosse. 2019. Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday*.

Duygu Cakir and Özge Yücel Kasap. 2020. Audio to video: Generating a talking fake agent. In *International Online Conference on Intelligent Decision Science*, pages 212–227. Springer.

Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. 2021. Optical flow based cnn for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, 146:31–37.

M Caldwell, JTA Andrews, T Tanay, and LD Griffin. 2020. Ai-enabled future crime. *Crime Science*, 9(1):1–13.

Kevin Matthe Caramancion. 2021. The demographic profile most at risk of being disinformed. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–7. IEEE.

Weiling Chen, Chenyan Yang, Gibson Cheng, Yan Zhang, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. 2018. Exploiting behavioral differences to detect fake ne. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 879–884. IEEE.

Hongmei Chi, Udochi Maduakor, Richard Alo, and Eleason Williams. 2020. Integrating deepfake detection into cybersecurity curriculum. In *Proceedings of the Future Technologies Conference*, pages 588–598. Springer.

Justin D Cochran and Stuart A Napshin. 2021. Deepfakes: awareness, concerns, and platform accountability. *Cyberpsychology, Behavior, and Social Networking*, 24(3):164–172.

Jesús Pérez Dasilva, Koldobika Meso Ayerdi, Terese Mendiguren Galdospin, et al. 2021. Deepfakes on twitter: Which actors control their spread? *Media and Communication*, 9(1):301–312.

Viktoria Degtereva, Svetlana Gladkova, Olga Makarova, and Eduard Melkostupov. 2020. Forming a mechanism for preventing the violations in cyberspace at the time of digitalization: Common cyber threats and ways to escape them. In *Proceedings of the International Scientific Conference-Digital Transformation on Manufacturing, Infrastructure and Service*, pages 1–6.

Rebecca A Delfino. 2020. Pornographic deepfakes: The case for federal criminalization of revenge porn's next tragic act. *Actual Probs. Econ. & L.*, page 105.

Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 26(1):69–91.

Luciano Floridi. 2018. Artificial intelligence, deepfakes and a future of ectypes. *Philosophy & Technology*, 31(3):317–321.

Anne Pechenik Gieseke. 2020. " the new weapon of choice": Law's current inability to properly address deepfake pornography. *Vand. L. Rev.*, 73:1479.

Alexander Godulla, Christian P Hoffmann, and Daniel Seibert. 2021. Dealing with deepfakes–an interdisciplinary examination of the state of research and implications for communication studies. *SCM Studies in Communication and Media*, 10(1):72–96.

Chandell Gosse and Jacquelyn Burkell. 2020. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, 37(5):497–511.

Shane Greenstein. 2021. The economics of confrontational conversation. *IEEE Micro*, 41(2):86–88.

The Guardian. 2021. Mother charged with deepfake plot against daughter's cheerleading rivals.

Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2020. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667.

Jeffrey T Hancock and Jeremy N Bailenson. 2021. The social impact of deepfakes. *Cyberpsychology, behavior and social networking*, 24(3):149–152.

Craig A Harper, Dean Fido, and Dominic Petronzi. 2021. Delineating non-consensual sexual image offending: Towards an empirical approach. *Aggression and violent behavior*, page 101547.

Susan Hazan. 2020. Deep fake and cultural truth-custodians of cultural heritage in the age of a digital reproduction. In *International Conference on Human-Computer Interaction*, pages 65–80. Springer.

Mark K Hinders and Spencer L Kirn. 2020. Cranks and charlatans and deepfakes. In *Intelligent Feature Selection for Machine Learning Using the Dynamic Wavelet Fingerprint*, pages 297–346. Springer.

Yoori Hwang, Ji Youn Ryu, and Se-Hoon Jeong. 2021. Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, 24(3):188–193.

Serena Iacobucci, Roberta De Cicco, Francesca Michetti, Riccardo Palumbo, and Stefano Pagliaro. 2021. Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, Behavior, and Social Networking*, 24(3):194–202.

Nanna Inie, Jeanette Falk Olesen, and Leon Derczynski. 2020. The rumour mill: Making the spread of misinformation explicit and tangible. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–4.

Nektaria Kaloudi and Jingyue Li. 2020. The ai-based cyber threat landscape: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–34.

Ignas Kalpokas. 2021. Problematising reality: the promises and perils of synthetic media. *SN Social Sciences*, 1(1):1–11.

Vasileia Karasavva and Aalia Noorbhai. 2021. The real threat of deepfake pornography: a review of canadian policy. *Cyberpsychology, Behavior, and Social Networking*, 24(3):203–209.

Catherine Kerner and Mathias Risse. 2021. Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics*, 8(1):81–108.

Ali Khodabakhsh, Raghavendra Ramachandra, and Christoph Busch. 2019. Subjective evaluation of media consumer vulnerability to fake audiovisual content. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020a. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146.

Jan Kietzmann, Adam J Mills, and Kirk Plangger. 2020b. Deepfakes: perspectives on the future "reality" of advertising and branding. *International Journal of Advertising*, pages 1–13.

Nils Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice–people cannot detect deepfakes but think they can. *Available at SSRN 3832978*.

Andrei OJ Kwok and Sharon GM Koh. 2021. Deepfake: A social construction of technology perspective. *Current Issues in Tourism*, 24(13):1798–1802.

Kevin LaGrandeur. 2021. How safe is our reliance on ai, and should we regulate it? *AI and Ethics*, 1(2):93–99.

Jack Langa. 2021. Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, 101:761.

Johannes Langguth, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková, and Daniel Thilo Schroeder. 2021. Don't trust your eyes: Image manipulation in the age of deepfakes. *Frontiers in Communication*, 6:26.

YoungAh Lee, Kuo-Ting Huang, Robin Blom, Rebecca Schriner, and Carl A Ciccarelli. 2021. To believe or not to believe: framing analysis of content and audience response of top 10 deepfake videos on youtube. *Cyberpsychology, Behavior, and Social Networking*, 24(3):153–158.

Sophie Maddocks. 2020. 'a deepfake porn plot intended to silence me': exploring continuities between pornographic and 'political'deep fakes. *Porn Studies*, 7(4):415–423.

Artem A Maksutov, Viacheslav O Morozov, Aleksander A Lavrenov, and Alexander S Smirnov. 2020. Methods of deepfake detection based on machine learning. In *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 408–411. IEEE.

Karolina Mania. 2020. The legal implications and remedies concerning revenge porn and fake porn: A common law perspective. *Sexuality & Culture*, 24(6):2079–2097.

David Moher, Larissa Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. 2015. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, 4(1):1–9.

Gillian Murphy and Emma Flynn. 2021. Deepfake false memories. *Memory*, pages 1–13.

Nicholas O'Donnell. 2021. Have we no decency? section 230 and the liability of social media companies for deepfake videos. *U. Ill. L. Rev.*, page 701.

Axel Oehmichen, Kevin Hua, Julio Amador Díaz López, Miguel Molina-Solana, Juan Gomez-Romero, and Yi-ke Guo. 2019. Not all lies are equal. a study into the engineering of political misinformation in the 2016 us presidential election. *IEEE Access*, 7:126305–126314.

Carl Öhman. 2019. Introducing the pervert's dilemma: a contribution to the critique of deepfake pornography. *Ethics and Information Technology*, pages 1–8.

Konstantin A Pantserev. 2020. The malicious use of ai-based deepfake technology as the new threat to psychological security and political stability. In *Cyber defence in the age of AI, smart societies and augmented humanity*, pages 37–55. Springer, Cham.

Miroslava Pavlíková, Barbora Šenkỳ˘rová, and Jakub Drmola. 2021. Propaganda and disinformation go online. *Challenging Online Propaganda and Disinformation in the 21st Century*, pages 43–74.

Adnan Qayyum, Junaid Qadir, Muhammad Umar Janjua, and Falak Sher. 2019. Using blockchain to rein in the new post-truth world and check the spread of fake news. *IT Professional*, 21(4):16–24.

Md Shohel Rana and Andrew H Sung. 2020. Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 70–75. IEEE.

Claudia Ratner. 2021. When "sweetie" is not so sweet: Artificial intelligence and its implications for child pornography. *Family Court Review*, 59(2):386–401.

Colonel Prof Filofteia Repez and Maria-Magdalena Popescu. 2020. Social media and the threats against human security deepfake and fake news. *Romanian Military Thinking*, (4).

Jean-Marc Rickli and Marcello Ienca. 2021. The security and military implications of neurotechnology and artificial intelligence. *Clinical Neurotechnology Meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*, page 197.

Helen Ronsner. 2021. The ethics of a deepfake anthony bourdain voice.

Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2019. Poster: Towards robust open-world detection of deepfakes. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2613–2615.

Eliza Strickland. 2018. Ai-human partnerships tackle" fake news": Machine learning can get you only so far-then human judgment is required-[news]. *IEEE Spectrum*, 55(9):12–13.

Eliza Strickland. 2019. Facebook takes on deepfakes. *IEEE Spectrum*, 57(1):40–57.

Catherine Stupp. 2019. Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. *The Wall Street Journal*, 30(08).

Bryan C Taylor. 2021. Defending the state from digital deceit: the reflexive securitization of deepfake. *Critical Studies in Media Communication*, 38(1):1–17.

Japan Times. 2020. Two men arrested over deepfake pornography videos.

Karin Wahl-Jorgensen and Matt Carlson. 2021. Conjecturing fearful futures: Journalistic discourses on deepfakes. *Journalism Practice*, pages 1–18.

Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11).

Jeffrey Westling. 2019. Are deep fakes a shallow concern? a critical analysis of the likely societal reaction to deep fakes. *A Critical Analysis of the Likely Societal Reaction to Deep Fakes (July 24, 2019)*.

Digvijay Yadav and Sakina Salmani. 2019. Deepfake: A survey on facial forgery technique using generative adversarial network. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 852–857. IEEE.

Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–14.

G Pascal Zachary. 2020. Digital manipulation and the future of electoral democracy in the us. *IEEE Transactions on Technology and Society*, 1(2):104–112.

Zsolt Ziegler. 2021. Michael polányi's fiduciary program against fake news and deepfake in the digital age. *AI & SOCIETY*, pages 1–9.

Human-in-the-Loop Systems for Truthfulness: A Study of Human and Machine Confidence

Yunke Qu
The University of Queensland
Brisbane, Australia
yunke.qu@uq.net.au

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Gianluca Demartini
The University of Queensland
Brisbane, Australia
demartini@acm.org

Abstract

Automatically detecting online misinformation at scale is a challenging and interdisciplinary problem. Deciding what is to be considered truthful information is sometimes controversial and difficult also for educated experts. As the scale of the problem increases, human-in-the-loop approaches to truthfulness that combine both the scalability of machine learning (ML) and the accuracy of human contributions have been considered.

In this work we look at the potential to automatically combine machine-based systems with human-based systems. The former exploit supervised ML approaches; the latter involve either crowd workers (i.e., human non-experts) or human experts. Since both ML and crowdsourcing approaches can produce a score indicating the level of confidence on their truthfulness judgments (either algorithmic or self-reported, respectively), we address the question of whether it is feasible to make use of such confidence scores to effectively and efficiently combine three approaches: (i) machine-based methods; (ii) crowd workers, and (iii) human experts. The three approaches differ significantly as they range from available, cheap, fast, scalable, but less accurate to scarce, expensive, slow, not scalable, but highly accurate.

1 Introduction

The challenge of identifying online misinformation has been rapidly growing given the increase in popularity of online news consumption as well as the ability to profile and micro-target social media users. Fighting the spread of online misinformation is a multi-disciplinary issue which requires both technical advances to process large amounts of false digital information as well as to understand the societal context in which such spreads happen. In order to best deal with the need to both scale to large number of fact-checks and have expert journalists manually checking and evaluating the veracity of posted information, human-in-the-loop systems have been considered (Demartini et al., 2020; Allen et al., 2021; Nakov et al., 2021).

Human-in-the-loop information systems aim at leveraging the ability of machines to scale and deal with very large amounts of data while relying on human intelligence to perform very complex tasks—for example, natural language understanding—or to incorporate fairness and/or explainability properties into the hybrid system (Demartini et al., 2017). Example of successful human-in-the-loop methods include ZenCrowd (Demartini et al., 2012), CrowdQ (Demartini et al., 2013), CrowdDB (Franklin et al., 2011), and Crowdmap (Sarasua et al., 2012). Active learning methods (Settles, 2009) are another example where labels are collected from humans, fed back to a supervised learning model, and then used to decide which data items humans should label next. Related to this is the idea of interactive machine learning (ML) (Amershi et al., 2014) where labels are automatically obtained from user interaction behaviors (Joachims and Radlinski, 2007).

While being more powerful than pure machine-based methods, human-in-the-loop systems need to deal with additional challenges to perform effectively and to produce valid results. One such challenge is the possible *noise* in the labels provided by non-expert humans. Depending on which human participants are providing labels, the level of data quality may vary. For example, making use of crowdsourcing to collect human labels from people online either using paid micro-task platforms like Amazon MTurk (Gadiraju et al., 2015) or by means of alternative incentives like, e.g., 'games with a purpose' (Von Ahn, 2006) is in general different from relying on a few experts.

There is often a trade-off between the cost and

Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 40–49,
Virtual, October 16-17, 2020.

the quality of the collected labels. On the one hand, it may be possible to collect few high-quality curated labels that have been generated by domain experts, while, on the other hand, it may be possible to collect very large amounts of human-generated labels that might not be 100% accurate. Since the number of available experts is usually limited, to obtain both high volume and quality labels, the development of effective quality control mechanisms for crowdsourcing is needed. Crowdsourcing as a method to collect labels to train veracity classification systems has recently been investigated (Roitero et al., 2020a,b; Soprano et al., 2021; Roitero et al., 2021).

Rather than seeing these data collection approaches as mutually exclusive, in this paper we focus on the possibility of combining machine-based truthfulness classifiers, non-expert annotators, and experts. In particular, we focus on the notion of *confidence*, i.e., the estimate of the reliability of the prediction—given by either a machine or a human annotator.

More in detail, in this paper we focus on the following research questions:

- RQ1: Can algorithmic and self-reported human confidence scores be used to reliably estimate the quality of truthfulness decisions?
- RQ2: Do humans and machines make similar or different mistakes in classifying truthfulness?
- RQ3: Can scarce expert annotator resources be integrated in such human-in-the-loop systems to intervene in cases when both crowd workers and machine-based truthfulness classifiers fail to correctly label an item?

To the best of our knowledge, this is the first attempt to understand the relationship between the effectiveness and confidence of the set including machine-based methods, crowd workers, and experts in a truthfulness classification task.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 details the methodology used in our study. We report and analyze our results in Section 4. Section 5 concludes by summarizing our findings and describing future work.

2 Related Work

In this section we summarize approaches computing and making use of confidence scores generated by ML models or human annotators (either self-reported or implicit).

Different types of ML methods are able to produce not only a classification decision, but to also attach a score that indicates how confident the algorithm is about the made decision. This is possible for a diverse set of methods, from decision trees to deep learning.

Poggi et al. (2017) consider a complete overview of 76 state-of-the-art confidence measures for ML; Mandelbaum and Weinshall (2017) discuss distance based confidence scores in the case of neural network based classifiers; Guo et al. (2017) detail a methodology to correctly interpret and compute confidence scores from ML models.

Trusting classification decisions solely based on algorithmic confidence may be risky. Once manually labelled data has been collected, trained models may reflect existing bias in the data. An example of such a problem is that of 'unknown unknowns' (UUs) (Attenberg et al., 2015), that is, data points for which a supervised model makes a high-confidence classification decision, which is however wrong. This means that the model is not aware of making mistakes. UUs are often difficult to identify because of the high-confidence of the model in its classification decision and may create critical issues in ML.

Quantifying decision confidence can also be done when decisions are made by human annotators. Hertwig (2012) discuss the role of confidence in the "wisdom of the crowd" paradigm. They point out how human confidence may be influenced by social interaction and the presence of others' annotations. Joglekar et al. (2013) describes methods to generate confidence intervals in order to capture crowd workers' confidence and bound accuracy scores. Jarrett et al. (2015) consider workers' self-assessment and investigates whether workers confidence correlates with quality and observe that self-evaluation is not indicative of their actual performance. This is consistent with findings by Gadiraju et al. (2017). Related to this observation, Li and Varshney (2017) show that workers annotation performance does not increase when considering the confidence scores to weight their contribution. Song et al. (2018) consider worker confidence in the setting of a labeling task performed with active learning techniques. Difallah et al. (2016) look at how to schedule labeling tasks to optimize their execution efficiency.

More than just human self-reported confidence, it is possible to implicitly measure confidence by, for example, computing inter-assessor agreement metrics. Nowak and Rüger (2010) study inter-annotator agreement and show how annotation quality can be improved when considering agreement scores to aggregate labels. Aroyo and Welty (2013) study the relationships between gold questions and workers agreement stating that agreement metrics do not necessary correlate with quality but may uncover alternative views on possible way to label data. Checco et al. (2017) discuss agreement measures applied to crowdsourcing and propose an alternative measure that is able to deal with sparse and incomplete data. Maddalena et al. (2017) incorporate assessor agreement into information retrieval evaluation metrics. In our work we make use of inter-annotator agreement metrics as a measure of human annotator confidence and quality.

3 Methodology

3.1 Dataset

We make use of manual truthfulness labels obtained from a crowdsourcing experiment as presented by Soprano et al. (2021). The crowdsourcing task was performed as follows. After an initial background survey phase, crowd workers are presented with 11 political statements, one after the other; 6 statements are taken from PolitiFact (Wang, 2017), 3 from ABC,[1] and 2 are used as quality checks. For each statement, according to the design defined by Roitero et al. (2020a), workers are asked to provide a truthfulness label. Additionally to the design by (Roitero et al., 2020a), we ask workers to also provide a confidence score on the expressed truthfulness label on a Likert scale in the $[-2, 2]$ range. The dataset contains a total of 120 statements from PolitiFact: 10 for each of the two political parties and for each level of the six-level truthfulness scale used by the expert assessors to evaluate the statements, and a total of 60 statements from ABC: 10 for each of the two political parties and for each level of the three-level truthfulness scale used by the expert assessors to evaluate the statements.

3.2 Machine Learning for Truthfulness Classification

BERT (Bidirectional Encoder Representations from Transformers) (Vaswani et al., 2017) is a language representation model based on performing a bidirectional training of a transformer based model. The core part of the model is the encoder / decoder architecture (Devlin et al., 2019), which is formed by different steps: the tokenization and numericalization of the input sequence followed by a set of embedding layers, which learn during the training phase a multidimensional embedding for each input token. Then, the learned representation is enriched with the context information represented with the positional encoding of the tokens built using the Multi Head (Self) Attention mechanism, which is fundamental to learn a better language model. In the BERT architecture multiple encoder / decoder blocks are stacked together to form the model. This architecture allows BERT to encode the entire input sequence at once, and perform two training task simultaneously: Masked Language Model and Next Sentence Prediction. The truthfulness classification task has been carried out using the BERT model pre-trained for classification tasks (`bert-base-uncased`[2]) fine-tuned with expert truthfulness labels on political statements. We use the output of the last softmax layer as the ML classification confidence score we use in our analysis.

GloVe (Global Vectors for Word Representation) by (Pennington et al., 2014) is a word vector learning technique which produces a vector space model similar to word2vec. The fundamental idea behind GloVe and word2vec is to learn, given a large corpus, a set of tuples containing a word and its context; then, the model is trained to predict the context given the specific word. Unlike word2vec which captures only the local context of a word, GloVe considers also the global context, implemented through a co-occurrence matrix. A feed-forward architecture with two dense layers (6 and 1 node, respectively), and a soft-max layer at the end. In Section 4 we only report results obtained with BERT for space constraints but results obtained with GloVe were similar.

[1] https://apo.org.au/collection/302996/rmit-abc-fact-check

[2] https://huggingface.co/bert-base-uncased

3.3 Crowdsourcing for Truthfulness Classification

With the crowdsourcing task design presented in Section 3.1, we collect non-expert labels from Amazon MTurk for 180 statements across different ground-truth truthfulness levels and different sources. In order to compare against supervised binary ML classifiers, we binarize human labels (originally collected on a 5-point $[-2, 2]$ Likert scale) by considering $\{-2, -1\}$ as the `False Statements` class and $\{1, 2\}$ as the `True Statements` class. We also binarize the 6-level Politifact scale and the 3-level ABC scale expert labels.

We use both crowd labels aggregated by the sum of the scores given by the 10 different workers who judged the same statement, as well as using the raw labels and confidence scores provided by individual crowd workers. We remove both the 20 ABC labels with an in-between value and the 5 aggregated crowd labels with a 0 value, as they do not indicate a binary classification decision. We are then left with 159 statements which we use in our analysis.

Thus, we generated a dataset that contains, for a total of 159 statements, truthfulness labels produced by ML models, non-expert crowd workers, and experts (i.e., ground truth labels) together with the respective confidence scores (experts are assumed to have max confidence).

3.4 ML and Crowd Confidence

To compute the crowd and machine learning confidence, we proceed as follows. For crowdsourced labels, we consider both the confidence scores self-reported by individual crowd workers, as well as the standard deviation among the ten crowd labels collected for each document. We refer these two scores respectively as *explicit* and *implicit* confidence scores.

Concerning the machine learning approaches, we cannot directly use the scores returned by the model in their last soft-max layer. Such scores can not be treated as confidence scores as shown in previous studies (Guo et al., 2017). Thus, to compute the machine learning confidence scores, we employed the bootstrap technique (Efron and Tibshirani, 1985): starting from a specific machine learning model, we produced ten different variations of such model obtained by varying the random seeds used in the initialization procedure; then, we run the ten models on the dataset and, similarly to what we do for crowdsourced labels, we compute the standard deviation over the ten scores collected for each document.

4 Results

4.1 ML and Crowd Accuracy

First we report on the truthfulness classification accuracy of both ML and crowd-based methods to label the truthfulness of statements in the dataset. As compared to expert ground-truth labels, ML models and crowd workers (with truthfulness labels for a statement aggregated by means of sum as raw labels are in $[-2, 2]$) perform at a similar level of accuracy (GloVe: 64.5%; BERT: 63.52%; word2vec: 62.9%; crowd: 55.3%). Thus, in the following we only report the results obtained on the most effective ML model.

Next, we explore the opportunity of combining these approaches for truthfulness classification by leveraging confidence-based combinations as well as involving scarce expert annotator resources when most beneficial.

4.2 ML and Crowd Confidence

Figure 1 shows both the ML (i.e., GloVe) and crowd confidence for the non-aggregated labels with a breakdown on the correctly and not correctly classified statements. Note that the ML and crowd confidence scores are shown in two separate plots since they are on two separate and not comparable scales: ML confidence scores are obtained from the bootstrap techniques applied to the soft-max layer of the ML algorithm which returns values in the $[0.5, 1]$ range, while the crowd confidence score is self-reported by each crowd worker on a [-2,2] scale. As we can see from Figure 1, ML confidence scores are almost always slightly lower on average for statements in which ML decisions are wrong and higher when ML correctly classify them (i.e., easy statements), even if such differences are small and not statistically significant. We see that crowd confidence shows the same behavior. Thus, answering **RQ1**, it seems raw confidence scores may be a weak signal indicating accurate classification decisions, thus leading to risks of undetectable classification errors (i.e., *unknown unknowns*) especially for the case of non-expert human annotators.

We now look at the confidence scores for the aggregated crowd labels; these confidence scores

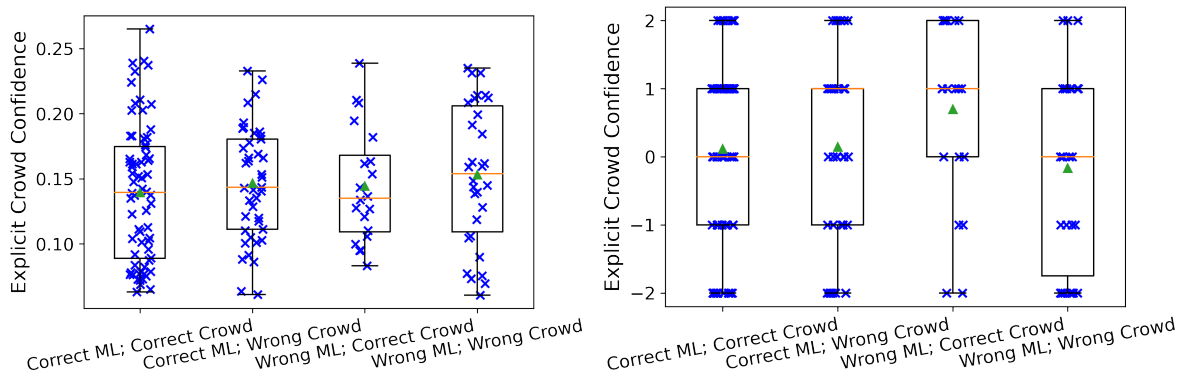


Figure 1: ML and explicit crowd confidence scores for raw crowd labels over correct and incorrect truthfulness classifications.

Figure 2: ML (left) and crowd confidence; both explicit (center plot) and implicit (right plot) for aggregated labels over ground-truth classes.

are obtained by taking the average value for each statement over all the workers who assessed it. Figure 2 shows, similarly to Figure 1 but with a breakdown on statement truthfulness rather than the correctness of its classification, the confidence for both ML and crowd truthfulness classification decisions.

As we can see from the plots, the mean confidence score for the 'true' statements is higher (although not significantly different according to a Mann-Whitney test) than the confidence score on the 'false' statements for confidence scores; on the contrary, for ML confidence scores the aggregated confidence scores are slightly higher (although not significantly different either) for the 'false' statements. This indicates that, similarly to what was observed for Figure 1, it seems that aggregated confidence scores are a weak signal indicating accurate classification decisions, and it should not be used as it may lead to undetectable classification errors.

We now move to study the relationship between ML and aggregated crowd confidence scores, to see if they are correlated and if one confidence score can act as a proxy for the other. Figure 3 shows on the x-axis the aggregated crowd confidence scores, on the y-axis the ML confidence; each dot is a statement; the different colors in the plot highlight a breakdown on either correctly and incorrectly classified statements by both the ML and the crowd. As we can see by inspecting the plots as a whole, both implicit and explicit crowd confidence show the same behavior when compared to ML confidence. Moreover, as we can see from inspecting the plots individually, the confidence scores for the statements correctly classified by both human and machine methods are spread across the plot; this is a further confirmation that trusting both ML and crowd confidence scores can lead to classification errors. If we now focus on the top-right and bottom-left part of the plots, we see that it contains dots of different colors; this indicates that even when both methods have either a high (top-right) or low (bottom-left) confidence scores the accuracy is similar. Again, this is a further confirmation of phenomena observed so far which indicates that both ML and crowd confidence scores should not be trusted.

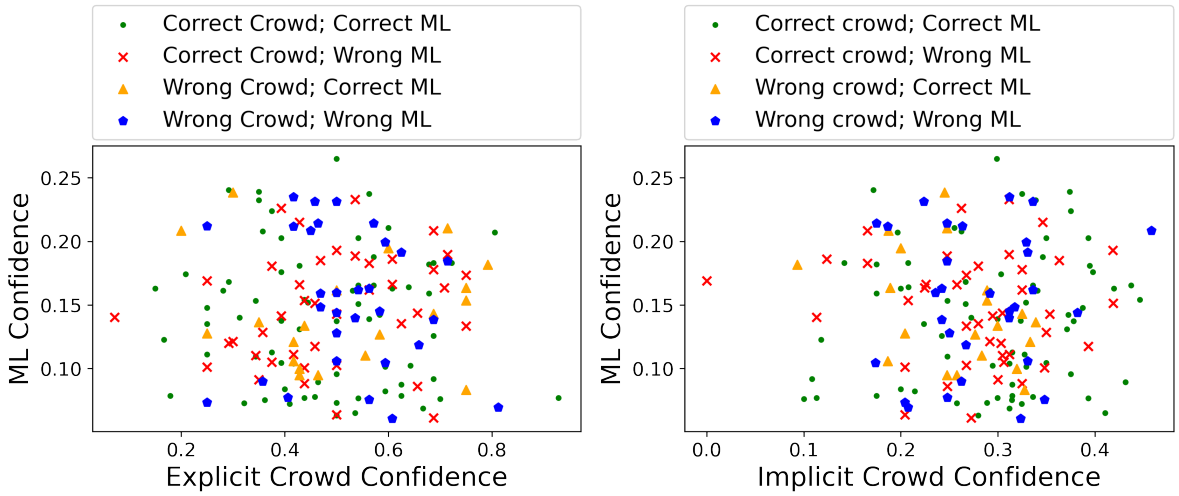


Figure 3: ML versus explicit (left plot) and implicit (right plot) crowd confidence with a breakdown on classification errors.

Summarizing the results observed so far, we can conclude that both ML and crowd confidence scores should be inspected carefully and not blindly trusted, as they can lead to classification errors. Furthermore, we observed a peculiar but interesting behavior for crowd confidence scores; both explicit (i.e., the scores submitted by the workers) and implicit (i.e., the ones automatically derived by considering the standard deviation of the truthfulness labels as submitted by the workers) confidence scores show a very similar behavior when compared to ML confidence scores; thus, this set of preliminary results hints that implicit confidence scores can act as a proxy for explicit scores if the aim is to compare them with ML scores. Thus, researchers and practitioners can avoid asking for explicit confidence scores if their focus is on accuracy and comparison with ML confidence scores, reducing the effort required by the crowd workers when performing the task.

To verify if this conjecture holds in general, we compared the explicit and implicit crowd confidence scores. Similarly to Figure 3, Figure 4 shows on the x-axis the aggregated crowd implicit confidence scores, and on the y-axis the aggregated crowd explicit confidence scores; each dot is a statement; the different colors in the plot highlight a breakdown on either correctly and incorrectly classified statements. As we can see from the plot, while implicit and explicit crowd confidence scores show a very similar behavior when compared to ML confidence (see Figure 3), we can see that the two measures are not correlated,

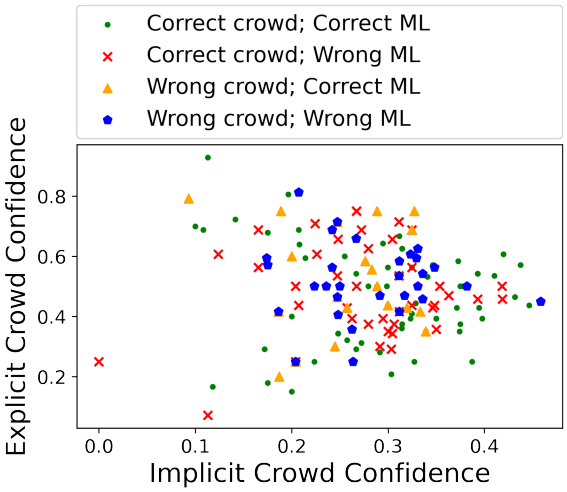


Figure 4: explicit versus implicit crowd confidence with a breakdown on classification errors.

and each statement shows a different implicit and explicit scores. Thus, if the focus of research and practitioners is purely on crowd confidence scores, implicit and explicit ones are substantially different. In the following we will focus on the relationship between effectiveness and confidence of the models, to investigate which crowd confidence scores provide a more informative signal when related to effectiveness.

We now turn to investigate whether the confidence and effectiveness of the methods used to predict the truthfulness of the statements are related. To this aim, we break down the confidence scores into quartiles and for each quartile we plot the accuracy of the considered method. Figure 5

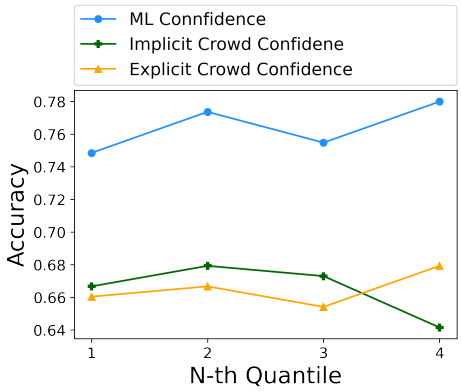


Figure 5: Confidence versus accuracy: group statements by quartiles of confidence scores and plot 4 points; both for ML and crowd.

shows the results, by displaying in the x-axis the confidence quartile, and in the y-axis the corresponding accuracy score; each series represent either the ML or crowd effectiveness scores. As we can see from the plot, there is no apparent clear pattern for all the series, even though it appears that the ML effectiveness scores overall observe a slight increase as the confidence scores itself increases, while the crowd scores, and in particular the implicit ones, observe a slight accuracy decrease while confidence increases.

Answering **RQ2**, we can see from the plots in Figure 3 and focusing on the yellow and blue statements, that there are many statements for which one of the two methods (i.e., ML or crowd) results in correct classification decisions, but the other method does not. Furthermore, Figure 5 shows that there is no clear signal that an increase in confidence is related to an increase in accuracy scores, for both ML or crowd.

While this negative results hint that it appears challenging to make use of confidence scores to increase the effectiveness of such methods and identify the cases where one of the two methods (i.e., ML or crowd) results in correct classification decisions but the other method does not, this set of results suggests the opportunity to investigate those signals in order to build an effective human-in-the-loop system which combines non-expert human and machine truthfulness classification together to obtain better quality decisions. We will discuss such approach in the following.

4.3 Can Confidence Be Leveraged?

Having studied the signal provided by both the ML and crowd confidence scores, we now investigate if such signals can be leveraged to improve the classification accuracy and the label quality when assessing the truthfulness of statements.

To this aim, and to answer **RQ3** about the potential involvement of experts, we perform the experiment as detailed in the following. Starting from the original dataset, for both ML and crowd, we replace the labels (i.e., the classification decisions for statements) that have the lower confidence scores with their corresponding ground truth label (i.e., the label as provided by the experts, which we assume to be always correct). Then, we re-compute the effectiveness of either the ML or crowd approach, measured by accuracy. To ensure a fair comparison, we also report the effectiveness of two baselines to compare against: the replacement with the ground truth label for a random statement in the dataset (repeated 50 times to remove random fluctuations of the series), and the replacement of the statements according to an oracle, which always replaces the statement that lead to obtain the highest increase in effectiveness. While the former baseline represents the average random case, the latter represents the optimal replacement selection strategy.

Figure 6 shows in the x-axis the number of statements which have been replaced in the original dataset, and in the y-axis either the ML or crowd accuracy scores; the three series represent the oracle, the random choice, and our strategy based on replacing the statements according to their confidence scores, replacing the ones with lower confidence first. As we can see focusing on the plot on the left side of Figure 6, the ML effectiveness increases as the replacements are done by removing the statements with lower confidence; we can also see that such strategy is always on average as effective as the random selection strategy, or even worse for same data points; both series are far less effective than the oracle. This results suggests that ML confidence can not act as a proxy for effectiveness, and thus it can not be leveraged (at least not in a naive way) to increase the model accuracy. This is not a definitive result and it suggest that there is room for improvement and it can be seen as an opportunity to study and develop novel methods to leverage confidence scores with the aim of identifying mis-classified statements and improving the overall model effectiveness. We leave for future work the analysis of more sophisticated approaches based on confi-

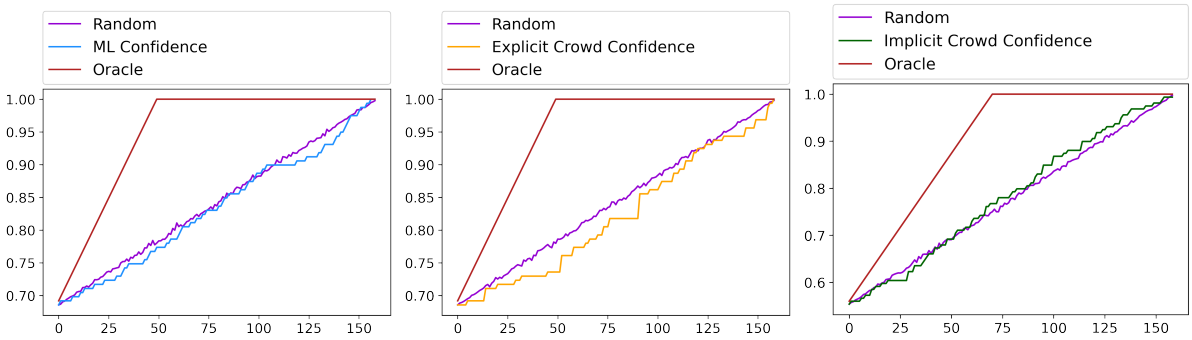


Figure 6: ML (left) and crowd (explicit, center; implicit, right) accuracy after replacing their labels with expert labels for statements (i) selected by an oracle (maximizing accuracy on each replacement), (ii) with lowest confidence, or (iii) uniformly at random.

dence or other signals. As we can see from the plot on the center of Figure 6, the same phenomena can be observed for crowd aggregated scores when explicit confidence scores are used. On the contrary, the situation changes when implicit confidence scores are used, as it can bee seen by inspecting the plot on the right side of Figure 6; such plot shows that, as the number of replacements grows, the accuracy of the methods grows and slightly over-performs the random replacement of statements. This is a positive result as it suggests that implicit confidence signals from crowd workers can be leveraged to increase the effectiveness of such method when employed to classify misinformation statements. These results are consistent with our previous observation on the lack of signal in ML confidence scores and that of previous work (Gadiraju et al., 2017; Li and Varshney, 2017) indicating that self-reported reliability is not accurate in crowdsourcing (i.e., highly confident crowd workers often make mistakes).

5 Conclusions and Future Work

In this paper we studied how ML and non-expert crowd workers classify the truthfulness of statements. To the best of our knowledge, this is the first attempt to study a human-in-the-loop pipeline for truthfulness classification which involves machines, non-experts (crowd workers), and experts (fact-checkers). In particular, we focused on both accuracy and confidence of the different approaches. We looked at both the accuracy and confidence signals alone, and we also studied their combination and their correlation; finally, we looked at identifying potential ways to leverage such signals and to combine them in order to improve the effectiveness of the classification decision process.

Our results show that, while ML and crowd confidence scores are not related to effectiveness, they can be leveraged to increase the effectiveness of the misinformation system. In this respect, implicit crowd confidence is a better indicator of effectiveness than crowd workers' self-reported confidence. We have also observed that ML and non-expert crowd workers make different mistakes, and their predictions do not agree in general. This result opens up to the opportunity of identifying more effective ways to combine these two approaches to increase the effectiveness of misinformation detection systems. Finally, we have shown that crowd workers and in particular their confidence scores can be leveraged to increase the effectiveness of systems when experts fact-checkers are brought into the loop in the cases where automatic ML or non-expert crowd workers are not confident on the submitted labels.

While our preliminary results are promising, there is still large room for improvement in making the most out of limited expert annotator resources; we believe this work is a first step towards the identification of signals for building an effective human-in-the-loop pipeline for misinformation assessment.

Acknowledgments. This work is supported by a Facebook Research award, the Australian Research Council (Grants No. DP190102141 and DE200100064), and by the ARC Training Centre for Information Resilience (Grant No. IC200100022).

References

Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Proceedings of WebSci*.

Joshua Attenberg, Panos Ipeirotis, and Foster Provost. 2015. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)*, 6(1):1–17.

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of HCOMP*, pages 11–20.

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of WWW*, pages 469–478.

Gianluca Demartini, Djellel Eddine Difallah, Ujwal Gadiraju, and Michele Catasta. 2017. An introduction to hybrid human-machine information systems. *Foundations and Trends in Web Science*, 7(1):1–87.

Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 43(3):65–74.

Gianluca Demartini, Beth Trushkowsky, Tim Kraska, Michael J Franklin, and UC Berkeley. 2013. CrowdQ: Crowdsourced query understanding. In *Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2016. Scheduling human intelligence tasks in multi-tenant crowd-powered systems. In *Proceedings of the 25th international conference on World Wide Web*, pages 855–865.

Bradley Efron and Robert Tibshirani. 1985. The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12(17):1–35.

Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. 2011. CrowdDB: Answering queries with crowdsourcing. In *Proceedings of SIGMOD*, pages 61–72.

Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems*, 30(4):81–85.

Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based preselection in crowdsourcing microtasks. *ACM Trans. Comput.-Hum. Interact.*, 24(4).

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*, pages 1321–1330.

Ralph Hertwig. 2012. Tapping into the wisdom of the crowd—with confidence. *Science*, 336(6079):303–304.

Julian Jarrett, Larissa Ferreira Da Silva, Laerte Mello, Sadallo Andere, Gustavo Cruz, and M Brian Blake. 2015. Self-generating a labor force for crowdsourcing: Is worker confidence a predictor of quality? In *Proceedings of the 2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, pages 85–90.

Thorsten Joachims and Filip Radlinski. 2007. Search engines that learn from implicit feedback. *Computer*, 40(8):34–40.

Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. 2013. Evaluating the crowd with confidence. In *Proceedings of KDD*, pages 686–694.

Qunwei Li and Pramod K Varshney. 2017. Does confidence reporting from the crowd benefit crowdsourcing performance? In *Proceedings of the 2nd International Workshop on Social Sensing (SocialSens)*, pages 49–54.

Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. 2017. Considering assessor agreement in ir evaluation. In *Proceedings of ICTIR*, page 75–82.

Amit Mandelbaum and Daphna Weinshall. 2017. Distance-based confidence score for neural network classifiers. *arXiv preprint arXiv:1709.09844*.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of IJCAI*, pages 4551–4558. Survey Track.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, pages 557–566.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. 2017. Quantitative evaluation of confidence measures in a machine learning world. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5228–5237.

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020a. Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background. In *Proceedings of SIGIR*, pages 439–448.

Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2021. Can the crowd judge truthfulness? a longitudinal study on recent misinformation about COVID-19. *Personal and Ubiquitous Computing*.

Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020b. The COVID-19 infodemic: Can the crowd judge recent misinformation objectively? In *Proceedings of CIKM*, pages 1305–1314.

Cristina Sarasua, Elena Simperl, and Natalya F Noy. 2012. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 525–541.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems*, 159:244–258.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management*, 58(6):102710.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Luis Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of ACL*, pages 422–426.

Cross-Lingual Rumour Stance Classification: a First Study with BERT and Machine Translation

Carolina Scarton and **Yue Li**
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, S1 4DP, Sheffield
{c.scarton, yli381}@sheffield.ac.uk

Abstract

Social media tend to be rife with rumours, which often have such high velocity and volume that fact-checkers struggle with debunking them with traditional methods. Prior research on English rumours has demonstrated that one can analyse the reactions (i.e. stance) expressed by social media users towards rumours, which ultimately enables automated flagging to journalists highly disputed rumours. This paper presents the first study of cross-lingual rumour stance classification. Through experiments with zero- and few-shot learning and in three languages (German, Danish and Russian), we show that models trained on English data can be used successfully for predicting stance in other languages. In the few-shot case, we also show that only few data points in the target language are needed to achieve the best results. In a multilingual setting, results for English are also further improved. Our results highlight the potential of multilingual BERT and machine translation for rumour analysis in languages where annotated data is scarce or not readily available.

1 Introduction

Social media are rife with rumours, which are fast-spreading, unverified pieces of information (Zubiaga et al., 2018). Journalists, however, are struggling to analyse misinformation in real time and at scale, thus motivating research into automatic rumour detection and analysis. A key rumour analysis task is *rumour stance classification* (RSC) (Li et al., 2019b; Dungs et al., 2018; Aker et al., 2019). RSC is useful, for instance, to flag rumours in a human-in-the-loop model, where fact-checkers could rely on crowd-based information, i.e. replies that support or deny a given rumour (Karmakharm et al., 2019). State-of-the-art automatic models for veracity prediction have also successfully used stance information in order to achieve best results. It is typically modelled as a four class problem, where posts replying to a rumour are classified as supporting; denying; questioning; or commenting on the rumour (Procter et al., 2013). In particular, the RumourEval 2017 (Derczynski et al., 2017) and 2019 (Gorrell et al., 2019) shared tasks demonstrated that RSC is a highly imbalanced problem, where the most informative classes, namely *support* and *deny*, are the minority classes.

Previous research on RSC, however, has focused predominantly on English, with the exception of Lozhnikov et al. (2018) for Russian and Lillie et al. (2019) for Danish. In the former, a small dataset (958 data points) composed of tweets and comments to media headlines is used to train RSC classifiers with word embeddings as features achieving 0.865 of macro-$F1$. In the latter, annotated Reddit posts in Danish (DAST) are used to train RSC classifiers using a varied feature set that includes word embeddings, part-of-speech, sentiment, and meta-data information. DAST has $3,007$ data points and the best model (SVM) achieves 0.421 of macro-$F1$.

This paper presents, to the best of our knowledge, the first study on cross-lingual RSC. We explore both multilingual BERT (MBERT) (Devlin et al., 2019) and machine translation (MT) approaches for zero-shot learning, and MBERT for few-shot learning as well as to train a full multilingual model. We make use of English, German, Russian and Danish RSC datasets, aiming to transfer the knowledge from English into other languages. Although our best results for Russian (macro$-F1 = 0.506$) and Danish (macro$-F1 = 0.352$) are not directly comparable to the performance of the respective language-specific models from prior work, we argue that cross-lingual RSC is a harder problem and that, given the small number of data points, previous work may be suf-

Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 50–59,
Virtual, October 16-17, 2020.

fering from overfitting. Moreover, the models trained in previous work require reliable Natural Language Processing tools (e.g. part-of-speech taggers, sentiment analysers) readily available in languages other than English, which is not usually the case. Nevertheless, ours is the first study to apply established multilingual models to perform cross-lingual RSC and thus enable RSC in low-resourced languages.

2 Related Work

In this section we discuss work related to RSC, focusing on the RumourEval datasets. RumourEval is a shared task, organised as part of SemEval in 2017 and 2019, comprising two subtasks: (A) stance classification and (B) veracity prediction. In A, the stance classification is formulated as a four-class classification problem, where replies to a social media post (source) can support, deny, query or provide a comment to the source. Subtask B consists of predicting the veracity of a rumour in social media, based on the text and/or metadata features. Successful systems in task B have also used the result of RSC as features (Li et al., 2019a,b). RSC (as formulated in the RumourEval datasets) is different from traditional stance classification tasks, since it also proposes the *query* class, useful in the rumour identification scenario. In addition, as pointed out by Scarton et al. (2020), in RSC, classes have different importance, with *support* and *deny* being the most interesting class for the task. This is particularly useful for human-in-the-loop application, where knowing if a reply is denying or supporting a post is more informative than if the reply is a comment.

Previous research in this area mainly focuses on the special characteristics of the datasets. Kochkina et al. (2017) employ Long Short Term Memory networks (LSTM) to capture the sequential nature of tweet threads. Yang et al. (2019) propose an inference chain-based system that utilise the information of the whole conversation. Task-specific features are also designed to boost the classifier performance (Aker et al., 2017; Bahuleyan and Vechtomova, 2017; Ghanem et al., 2019). Dealing with skewed distribution towards the *comment* class is another significant direction as most of the systems suffer the low performance over the minority classes, especially the *deny* class. Li and Scarton (2020) compare the performance of traditional imbalanced data treatments on the RumourEval datasets. They design a simple BERT-based model combining with threshold-moving, ranking first and second in RumourEval 2017 and 2019 respectively.

However, RumourEval datasets only consider English posts. To the best of our knowledge, Lozhnikov et al. (2018) (for Russian) and Lillie et al. (2019) (for Danish) are the only previous work tackling RSC for languages other than English. For both, datasets are created following the same annotation scheme as RumourEval, i.e. replies to comments are annotated in one of the four classes (for more details about these datasets see Section 3.1). In (Lozhnikov et al., 2018), feature-based classifiers are trained, using pre-trained word embeddings for Russian. They achieve an impressive 0.865 of macro-$F1$, with great performance for *support* and *deny* classes. In the work for Danish, they experiment with a LSTM classifier and several feature-based models, using the DAST dataset. Different feature types were explored, including textual information, sentiment, bag-of-words, part-of-speech, word frequency and information from the Reddit metadata. Their best model, achieving 0.421 of macro-$F1$, is an SVM trained with hyperparameter optimisation and feature selection (e.g. Reddit-based features are not included in this model). (Lillie et al., 2019) is also the first work to present cross-lingual veracity prediction. Veracity classifiers are trained relying on language independent information using the PHEME dataset (Zubiaga et al., 2016) for training and DAST for testing and vice-versa. Results suggest that cross-lingual models are comparable to monolingual models.

Although these two previous work represent an advance in RSC for languages other than English, the processing of collecting monolingual annotated data is expensive and time-consuming to be feasible for all languages. In addition, they assume that NLP resources, such as pre-trained word embeddings, part-of-speech taggers and sentiment analysis models, are readily available in the language under study, which is not a reality for most languages other than English. Therefore, it is important to explore approaches that enable low-resourced languages to benefit from the relatively large amount of English training data. Although cross-lingual approaches have been investigated for various NLP problems (Stappen et al., 2020; Chidambaram et al., 2019; Eriguchi et al., 2018),

	Support	Deny	Question	Comment
EN(training)	841 (19.8%)	333 (7.8%)	330 (7.8%)	2,734 (64.5%)
EN(test)	94 (9.0%)	71 (6.8%)	106 (10.1%)	778 (74.1%)
DE (test only)	48 (17.0%)	13 (4.6%)	18 (6.4%)	203 (72.0%)
DA(training)	184 (8.2%)	232 (10.4%)	61 (2.7%)	1756 (78.6%)
DA(test)	89 (11.5%)	68 (8.8%)	20 (2.6%)	597 (77.1%)
RU(training)	35 (5.0%)	36 (5.2%)	139 (20.1%)	481 (69.6%)
RU(test)	23 (8.6%)	10 (3.7%)	53 (19.9%)	181 (67.8%)

Table 1: Data distribution of classes in each dataset (values in parenthesis are the percentages of each class).

to the best of our knowledge, there is no research on cross-lingual RSC task.

3 Experimental Settings

3.1 Datasets

English The English model is trained on the RumourEval 2017 (RE2017) dataset (Derczynski et al., 2017) which has $4,238$ source-reply tweet pairs from eight different events in the training set: the Ferguson unrest, the shooting at Charlie Hebdo, the hostage situation in Sydney, the Germanwings plane crash, the Ottawa shooting, a rumour about a coup in Russian, a rumour that Prince was doing a surprise show in Toronto, and a rumour that Footballer Michael Essien had contracted Ebola. The test set has $1,049$ tweet pairs from ten events (the same eight events in the training data plus: a rumour that Hillary Clinton was diagnosed with pneumonia during the 2016 US elections and rumour that Youtuber Marina Joyce had been kidnapped).

German The German data (Zubiaga et al., 2016) has 282 tweet pairs from three different events: the Germanwings plane crash, a rumour about a coup in Russian, and a rumour about the Gurlitt collection.

Danish For Danish, we use the DAST dataset with $3,007$ source-reply Reddit pairs (Lillie et al., 2019). It encompasses posts from 11 rumourous events: 5G, Donald Trump, HPV vaccine, ISIS, *Kost* (diet), MeToo movement, *Overvågning* (surveillance), Peter Madsen, *Politik* (politics), *Togstrejke* (train strike), and *Ulve i DK* (wolves in Denmark).

Russian For Russian we use a dataset with source-reply tweet pairs concatenated with claim-reply pairs of Meduza[1] and Russian Today.[2] It has 958 pairs divided into 17 threads covering different topics (Lozhnikov et al., 2018).[3]

For monolingual and few-shot learning experiments, we divide the Danish and Russian datasets into training and test sets. For Danish, eight events are used for training (ISIS, *Kost*, MeToo, *Overvågning*, Peter Madsen, *Politik*, *Togstrejke*, and *Ulve i DK*) and three for testing (5G, HPV vaccine, and Donald Trump). For Russian, 14 topics are used for training and three for testing. Dividing the training and test sets using events/topics is expected to minimise the chances of overfitting, since, at training time, the models will not see the events that appear in the test set. The German dataset is rather small, with only one of the three events having data points in all classes, which makes it unsuitable for data splitting. Therefore, this dataset is only used as a test set in the zero-shot experiments. Table 1 shows the class distributions for each dataset.

3.2 Models

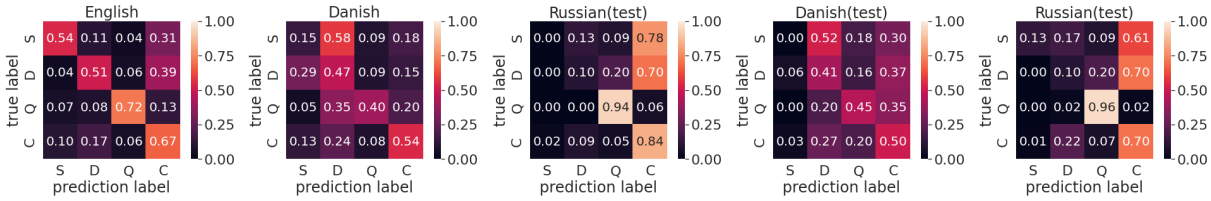
Settings BERT models are fine-tuned for three epochs with a batch size of 16, 12 transformer layers, hidden unit size of 768, 12 attention heads, and 110M parameters using the `ktrain` toolkit (Maiya, 2020). We apply the 1 *cycle policy* (Smith, 2018) for training and search the optimal learning rate among $5e^{-5}$, $3e^{-5}$, $1e^{-5}$, and $1e^{-4}$. For dealing with data imbalance, we follow (Li and Scarton, 2020) and apply threshold moving (TM) (Maloof, 2003; Sheng and Ling, 2006), where the classifier is trained with the imbalanced data, but the decision threshold that transforms the output probability into class labels is changed. We set the threshold according to the class proportions based on two assumptions: (1) the class proportion in the test set is similar to that of the training set; and (2) the prior of a class is equivalent to its pro-

[1] https://meduza.io/en
[2] https://www.rt.com

[3] More details about the topics are available at (Lozhnikov et al., 2018).

		macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
SOTA RE2017	EN(test)	0.452	0.363	0.296
MBERT_EN	EN(test)	**0.528**	**0.602**	**0.487**
MBERT_DA	DA(test)	**0.300**	**0.350**	**0.251**
MBERT_RU	RU(test)	0.442	0.000	0.211
MBERT_MTDA	DA(test)	0.228	0.00	0.166
MBERT_MTRU	RU(test)	**0.467**	**0.306**	**0.259**

Table 2: Results for the monolingual MBERT models (best results are shown in bold).



(a) MBERT_EN (b) MBERT_DA (c) MBERT_RU (d) MBERT_MTDA (e) MBERT_MTRU

Figure 1: Confusion matrices for monolingual models.

portion in the training set (Collell et al., 2018).

Multilingual BERT (MBERT) use the pre-trained *BERT-base-multilingual-cased* model.[4] The hypothesis of using a MBERT model trained only on English data (MBERT_EN) for other languages is that it would be capable of performing zero-shot RSC, similar to its success in other NLP tasks (Pires et al., 2019). We also experiment with models trained only on Danish (MBERT_DA) or Russian (MBERT_RU) training data. *Few-shot* learning is also explored, where MBERT_EN is further fine-tuned using the training data in Danish (MBERT_ENDA) or Russian (MBERT_ENRU). We aim to check whether monolingual data, even just a few data points, can help to improve the performance of SRC in Danish or Russian. Finally, we also propose a full multilingual model, where the English, Danish and Russian training sets are combined and used for training the model.

Machine Translation (MT) is used to translate the Russian, Danish and German data into English, so the English-only models can be applied. We use Google Translate[5] for producing the automatic translations and MBERT_EN to classify the translated text (MT+MBERT_EN model). We also use MT to translate the English training data into Russian or Danish, and fine-tune MBERT monolingual models (MBERT_MTDA and MBERT_MTRU models for Danish and Russian, respectively).

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://translate.google.co.uk

3.3 Evaluation

Scarton et al. (2020) show that the evaluation metrics used in RumourEval in 2017 (accuracy) and 2019 (macro-$F1$) are not robust for this four-class imbalanced classification task. They suggest the use of two alternative metrics: geometric mean of recall (GMR) and $wF2$. GMR heavily penalises models that underperform on minority classes, being an useful metric for imbalanced classification tasks. $wF2$ is a weighted version of macro-$F2$ that gives more importance to recall than precision and also assigns higher weights for the most important RSC classes, i.e. *support* and *deny*. Therefore, in this paper, besides reporting macro-$F1$ for comparison with previous work, we also report $wF2$ and GMR.[6]

4 Cross-lingual Rumour Stance Classification

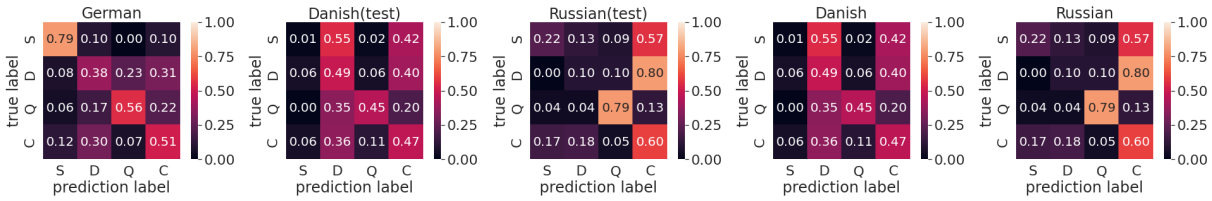
4.1 Monolingual models

Aiming to assess the effectiveness of zero- and few-shot models, we devise monolingual models for Danish and Russian, either using our pre-defined training sets (MBERT_DA and MBERT_RU) or the machine translated training sets (MBERT_MTDA and MBERT_MTRU). Results shown in Table 2 also include values for MBERT_EN in the RE2017 test set and for the best model in the RE2017 shared task (Best

[6] For $wF2$, we use the same weights as Scarton et al. (2020), i.e $w_{deny} = w_{support} = 0.40$, $w_{query} = 0.25$ and $w_{comment} = 0.05$

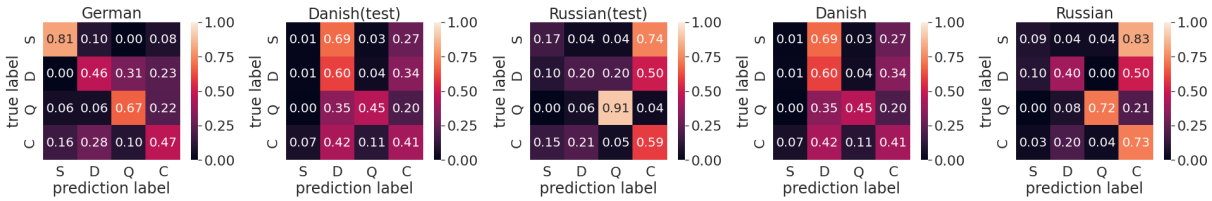
	MBERT_EN			MT+MBERT_EN		
	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
DE	**0.470**	0.542	0.480	0.464	**0.585**	**0.505**
DA(full)	**0.259**	0.221	0.201	0.248	**0.228**	**0.219**
DA(test)	**0.241**	0.184	0.187	0.234	**0.188**	**0.200**
RU(full)	**0.419**	**0.377**	**0.278**	0.406	0.360	0.260
RU(test)	0.420	0.319	0.252	**0.437**	**0.368**	**0.275**

Table 3: Results for zero-shot learning using MBERT_EN model or MT+MBERT_EN. Best values are in bold.



(a) (b) (c) (d) (e)

Figure 2: Confusion matrices for zero-shot learning using MBERT_EN.



(a) (b) (c) (d) (e)

Figure 3: Confusion matrices for zero-shot learning using MT+MBERT_EN.

RE2017), showing that MBERT_EN would have ranked first in this shared task.[7] Interestingly, MBERT_MTRU performs significantly better than MBERT_RU, which may be explained by the small size of the Russian training set. Conversely, MBERT_DA performs best, probably due to more in-domain data available.

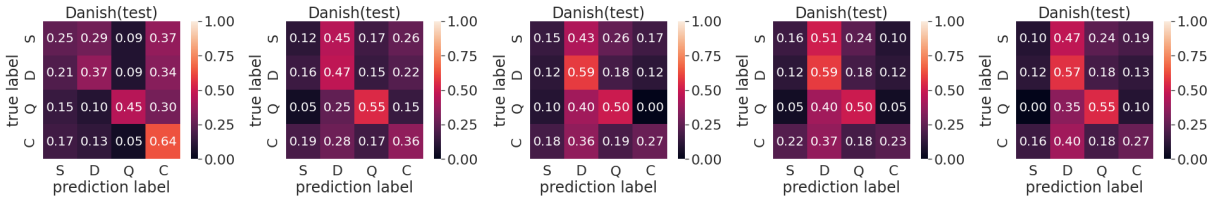
Figure 1 shows the confusion matrices for the monolingual models. MBERT_RU has $GMR = 0$ because it fails to predict all *support* instances. MBERT_DA also underperforms for the *support* class and has a bias towards *denies* (mainly by predicting *supports* as *denies*). MBERT_DA is the best for Danish, since it predicts 15% of *support* and 47% of *deny* correctly, versus 0% and 41% for *support* and *deny*, respectively, for MBERT_MTDA.

4.2 Zero-Shot Rumour Stance Classification

In the first zero-shot experiment, MBERT_EN model is used for RSC in other languages. We then compare this model with a pipelined approach using MT+MBERT_EN, where the data in Danish, Russian or German are machine translated into English and classified using MBERT_EN. Table 3 shows the results of evaluation in the Danish and Russian full and test sets and the German set, whilst Figure 2 and Figure 3 show the confusion matrices for models MBERT_EN and MT+MBERT_EN, respectively. Results for German are particularly good, being comparable to the results for English (Table 2). One reason for this high performance is that the German test set includes tweets about rumours that appear in the English training set.

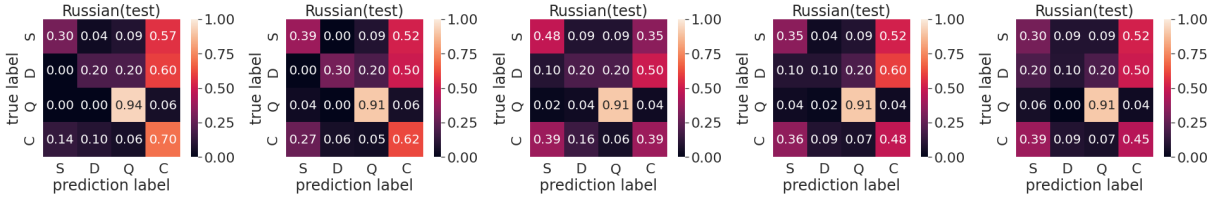
For Danish, the best GMR and $wF2$ are achieved with the pipelined MT+MBERT_EN model (for both full and test), however, these results are worse than the monolingual model (Table 2). The main issue is with the misclassification of *supports* (-0.14 of class accuracy in comparison to MBERT_DA in Figure 1b). Data characteristics may justify this low performance: while the English training data is composed of tweets, the Danish data has Reddit posts, which are consid-

[7] The best model for RE2017 according to the reported metrics is NileTMRG (Enayet and El-Beltagy, 2017). This differs from the winner of the task (Kochkina et al., 2017), which shows low scores for all metrics (Scarton et al., 2020).



(a) Full training set (b) $n = 244$ (c) $n = 100$ (d) $n = 60$ (e) $n = 20$

Figure 4: Confusion matrices for few-shot learning using MBERT_EN model as starting point for Danish (MBERT_ENDA).



(a) Full training set (b) $n = 140$ (c) $n = 100$ (d) $n = 60$ (e) $n = 20$

Figure 5: Confusion matrices for few-shot learning using MBERT_EN model as starting point for Russian (MBERT_ENRU).

erably longer and use different argumentation patterns.[8] For Russian, MT+MBERT_EN also shows best results for the test set with improvements due to better performance in the *supports* class (+0.17 in comparison to MBERT_RU in Figure 1c). On the other hand, the full set in Russian achieves best results when MTBERT_EN is applied.

For German (Figures 2a and 3a) and Danish (full: 2d and 3d; and test: 2b and 3b), the best results in GMR and $wF2$ for MT+BERT_EN is explained because this model outperforms BERT_EN for classes *support* and *deny*. For RU(full), MBERT_EN (Figure 2e) is significantly better at predicting *denies* than MT+MBERT_EN (Figure 3e). On the other hand, for RU(test) MT+BERT_EN (Figure 3c) shows significantly better results for *deny* and *query* classes than BERT_EN (Figure 2c).

4.3 Few-shot Rumour Stance Classification

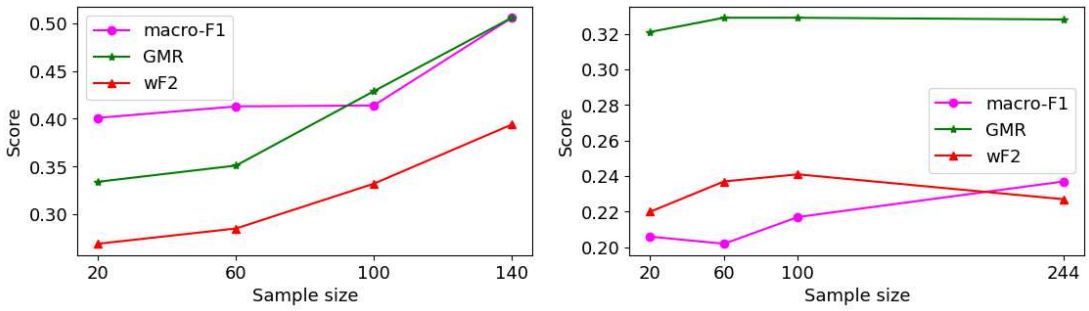
For few-shot learning, we use MBERT_EN as the starting point and continue fine-tuning it with the target language training data, i.e. either DA(training) or RU(training). When monolingual data is available for training, the hypothesis is that MBERT models would benefit from the pre-training on a larger dataset (English) and specialise their performance using target language data. Table 4 shows the performance of models trained in this setting: MBERT−ENDA for Danish (the confusion matrix is show in Figure 4a) and MBERT−ENRU for Russian (the confusion matrix is show in Figure 5a). Results for Russian show a significant increase in performance over the monolingual and zero-shot models, specially in terms of GMR and $wF2$. This happens mainly due to improvements in the accuracy of *supports* (+0.30) and *denies* (+0.10). Few-shot learning also improves the results for Danish, mainly because the MBERT_ENDA model better handles all classes (specially *support*, improving +0.10 points), without biasing towards *denies*.

	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
DA(test)	**0.352**	**0.401**	**0.295**
RU(test)	0.501	0.448	0.349
DA(balanced)	0.237	0.328	0.227
RU(balanced)	**0.506**	**0.506**	**0.394**

Table 4: Few-shot learning using MBERT_EN model as starting point (best results are shown in bold).

Balanced data re-sampling We under-sampled the Russian and Danish training sets, so that all classes have the same number of data points. For Russian, since the class with fewest examples (*support*) has 35 instances, 140 is the size of this balanced training set. For Danish, the smallest class is *query* with 61 examples, so the balanced set has 244 data points. Results for models trained

[8] We have also experimented with an MBERT_EN model trained on RumourEval 2019 data that contains Reddit posts. Results for Danish did not improve, probably due to the size of the English Reddit sample (only 16.9% of the training data).



(a) MBERT_ENRU (b) MBERT_ENDA

Figure 6: Performance of: (a) MBERT_ENRU varying the sample sizes of the Russian training set and MBERT_ENDA varying the sample sizes of the Danish training set.

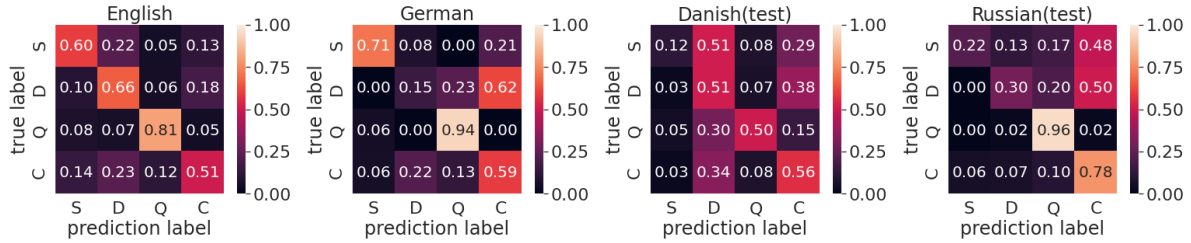


Figure 7: Confusion matrices for the full multilingual model.

on these balanced sets are shown in the bottom part of Table 4: DA(balanced) (see Figure 4b for the confusion matrix) and RU(balanced) (see Figure 5b for the confusion matrix). For Russian, this approach outperforms the use of the entire training set, with further improvements in the *support* (+0.09) and *deny* (0.10) classes. For Danish, the balanced approach is not better than using the entire training set.

Monolingual data sample size Aiming to assess the amount of monolingual data needed to outperform zero-shot learning, we also decrease the size of the samples gradually, starting from 35 (or 61) and stopping at 5 data points per class. For Russian (Figure 6a), 100 data points of balanced training data is enough to outperform zero-shot approaches ($GMR = 0.429$ and $wF2 = 0.332$). For Danish (Figure 6b), 20 data points is enough to improve over zero-shot learning, with $GMR = 0.303$ and $wF2 = 0.220$. The confusion matrices for this experiment are also shown: Figures 4b to 4e for Danish and Figures 5b to 5e for Russian. We observe that for Danish, the best performance for *support* class is achieved when the full dataset is used (Figure 4a), whilst the best performance for *denies* is reached when the samples size (n) is 60 or 100. In particular, there is a significant drop in the performance of *denies* when the data sample is increased to $n = 244$, which justifies the decrease in $wF2$ show in Figure 6b. The main issue with the balanced Danish models is the bias towards the *deny* class, which is minimised when the full Danish traning set is used. For Russian, the best performance for *support* is at $n = 100$, whilst the *deny* is more accurately predicted when $n = 140$.

4.4 Full multilingual model

To build a *full* multilingual model, we fine-tune MBERT with all training data in all languages. We aim to assess whether joint training improves performance in the individual languages. Table 5 shows the results for this experiment and Figure 7 shows the confusion matrices.

	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
EN	0.484	0.635	0.509
DA(test)	0.323	0.366	0.263
RU(test)	0.524	0.470	0.368
DE (zero-shot)	0.502	0.497	0.462

Table 5: Results for a MBERT fine-tuned with all training data for all languages.

Significant improvements are shown for English in terms of GMR and $wF2$, thanks to significant improvements in accuracy on the *support* (+0.06), *deny* (+0.15) and *query* (+0.09) classes. Even

though the macro-$F1$ for English here is worse than the monolingual model (Table 2), our multilingual model still outperforms the state-of-the-art for the RE2017 shared task. Results for Danish and Russian are better than zero-shot models, although worse than few-shot models. In few-shot learning, the models get more specialised in the target language, while in the multilingual setting the variety of data may harm the prediction for languages with fewer data points. For German, this is also a zero-shot setting, since we do not have training data for this language. Results in terms of GMR and $wF2$ are worse in this multilingual setting than in our zero-shot experiment (Table 3), mainly because the performance of *support* is significantly harmed. We hypothesise that the variety of data introduced by Danish and Russian can be harming the performance for German, that is a very similar set to the English training data.

5 Conclusions

To the best of our knowledge, this is the first paper to produce a detailed comparison of cross-lingual RSC on four languages (English, German, Danish, and Russian) and across different types of posts (tweets, Reddit posts, and comments to media headlines). The results of our zero-shot learning experiments show that both MT- and MBERT-based RSC can be useful for low-resourced languages, where no data is available for training.

Few-shot learning shows the best performance for both Danish and Russian, outperforming zero-shot models with just a few data points in the target language. Therefore, monolingual data can be useful for improving models, but only a few data points are actually needed (in our experiments, models outperforming the zero-shot experiments were achieved with 100 data points for Russian and 20 for Danish). A full multilingual model improved the performance for English, showing that data in other languages may also be helpful for high-resource languages.

We argue that cross-lingual RSC can also enable the analysis of trending rumours, that may have replies in multiple languages. In particular, MBERT-based approaches can also be useful for robustly model code-switching, where a single reply contains words in multiple languages. Future work include further experiments with more languages (given the availability of data) and the use of cross-lingual RSC for supporting the task of veracity prediction.

Acknowledgements

This work was funded by the WeVerify project (EU H2020, grant agreement: 825297).

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria. INCOMA Ltd.

Ahmet Aker, Alfred Sliwa, Fahim Dalvi, and Kalina Bontcheva. 2019. Rumour verification through recurring information and an inner-attention mechanism. *Online Social Networks and Media*, 13:100045.

Hareesh Bahuleyan and Olga Vechtomova. 2017. UWaterloo at SemEval-2017 task 8: Detecting stance towards rumours with topic independent features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 461–464, Vancouver, Canada. Association for Computational Linguistics.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

Guillem Collell, Drazen Prelec, and Kaustubh R. Patil. 2018. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, 275:330–340.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Omar Enayet and Samhaa R. El-Beltagy. 2017. NileTMRG at SemEval-2017 task 8: Determining rumour and veracity support for rumours on Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474, Vancouver, Canada. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Bilal Ghanem, Alessandra Teresa Cignarella, Cristina Bosco, Paolo Rosso, and Francisco Manuel Rangel Pardo. 2019. UPV-28-UNITO at SemEval-2019 task 7: Exploiting post's nesting and syntax information for rumor stance classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1125–1131, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 115–120, Hong Kong, China. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019a. eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 855–859, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019b. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Yue Li and Carolina Scarton. 2020. Revisiting rumour stance classification: Dealing with imbalanced data. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 38–44, Barcelona, Spain (Online). Association for Computational Linguistics.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221, Turku, Finland. Linköping University Electronic Press.

Nikita Lozhnikov, Derczynski Leon, and Mazzara Manuel. 2018. Stance Prediction for Russian: Data and Analysis. In *Proceedings of 6th International Conference in Software Engineering for Defence Applications*, pages 176–186, Roma, Italy. Advances in Intelligent Systems and Computing, Springer, Cham.

Arun S. Maiya. 2020. ktrain: A Low-Code Library for Augmented Machine Learning. *arXiv preprint arXiv:2004.10703*.

Marcus A Maloof. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of the ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, Washington, DC, USA.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, 16(3):197–214.

Carolina Scarton, Diego Silva, and Kalina Bontcheva. 2020. Measuring what counts: The case of rumour stance classification. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 925–932, Suzhou, China. Association for Computational Linguistics.

Victor S. Sheng and Charles X. Ling. 2006. Thresholding for making classifiers cost-sensitive. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, pages 476–481, Boston, Massachusetts. American Association for Artificial Intelligence.

Leslie N. Smith. 2018. *A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay*. US Naval Research Laboratory Technical Report 5510-026.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel. *arXiv preprint arXiv:2004.13850*.

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.

Author Index