AMUSED: An Annotation Framework of Multi-modal Social Media Data

Gautam Kishore Shahi
University of Duisburg-Essen, Germany
gautam.shahi@uni-due.de

Abstract

In this paper, we present a semi-automated framework called AMUSED for gathering multi-modal annotated data from the multiple social media platforms. The framework is designed to mitigate the issues of collecting and annotating social media data by cohesively combining machine and human in the data collection process. From a given list of the articles from professional news media or blog, AMUSED detects links to the social media posts from news articles and then downloads contents of the same post from the respective social media platform to gather details about that specific post. The framework is capable of fetching the annotated data from multiple platforms like Twitter, YouTube, Reddit. The framework aims to reduce the workload and problems behind the data annotation from the social media platforms. AMUSED can be applied in multiple application domains, as a use case, we have implemented the framework for collecting COVID-19 misinformation data from different social media platforms.

Introduction

With the growth of the number of users on different social media platforms, social media have become part of our lives. They play an essential role in making communication easier and accessible. People and organisations use social media for sharing and browsing the information, especially during the time of the pandemic, social media platforms get massive attention from users Talwar et al. (2019). Braun and Gillespie conducted a study to analyse the public discourse on social media platforms and news organisation. The design of social media platforms allows getting more attention from the users for sharing news or user-generated content. Several statistical or computational study has been conducted using social media data Braun and Gillespie (2011). But data gathering and its annotation are time-consuming and financially costly. In this study, we resolve the complications of data annotation from social media platforms for studying the problems of misinformation and hate speech.

Usually, researchers encounter several problems while conducting research using social media data, like data collection, data sampling, data annotation, quality of the data, and the bias in data Grant-Muller et al. (2014). Data annotation is the process of labelling the data available in various formats like text, video or images. Researchers annotate social media data for researches based on hate speech, misinformation, online mental health etc. For supervised machine learning, labelled data sets are required so that machine can quickly and clearly understand the input patterns. To build a supervised or semi-supervised model on social media data, researchers face two challenges- timely data collection and data annotation Shu et al. (2017). One time data collection is essential because some platforms either restrict data collection or often the post itself is deleted by social media platforms or by the user. For instance, Twitter allows data crawling of only the past seven days (from the date of data crawling) by using the standard APIs Stieglitz et al. (2018). Moreso, it is not possible to collect the deleted posts from social media platforms. Another problem stands with data annotation; it is conducted either in an in-house fashion (within lab or organisation) or by using a crowd-based tool(like Amazon Mechanical Turk(AMT)) Aroyo and Welty (2015). Both approaches of data annotations require an equitable amount of effort to write the annotation guidelines along with expert annotators. In the end, we are not able to get quality annotated data which makes it challenging to a reliable statistical or artificial intelligence based analysis. There is also always a chance of wrongly labelled data leading to bias in data Cook and Stevenson (2010).

Currently, professional news media or blogs also cover the posts from social media posts in their articles. The inclusion of social media posts in the news and blog articles creates an opportunity to gather labelled social media data. Journalists cover humongous topics of social issues such as misinformation, propaganda, rumours during elections, disasters, pandemics, and mob lynching, and other similar events. Journalists link social media posts in the content of the news articles or blogs to explain incidents Carlson (2016).

To solve the problems of on-time data collection and data annotation, we propose a semi-automatic framework for data annotation from social media platforms. The proposed framework is capable of getting annotated data on social issues like misinformation, hate speech or other critical social scenarios. The key contributions of the paper are listed below-

- We present a semi-automatic approach for gathering annotated data from social media platforms. AMUSED gathers labelled data from different social media platform in multiple formats(text, image, video).

- AMUSED reduces the workload, time and cost involved in traditional data annotation technique.
- AMUSED resolves the issues of bias in the data (wrong label assigned by annotator) because the data gathered will be labelled by professional news editors or journalists.
- The AMUSED can be applied in many domains like fake news or propaganda in the election, mob lynching etc. for which it is hard to gather the data.

To present a use case, we apply the proposed framework to gather data on COVID-19 misinformation on multiple social media platforms. In the following sections, we discuss the related work, different types of data circulated and its restrictions on social media platforms, current annotation techniques, proposed methodology and possible application domain; then we discuss the implementation and result. We also highlight some of the findings in the discussion, and finally, we discuss the conclusion and ideas for future works.

Related Work

Much research has been published using social media data, but they are limited to a few social media platforms or language in a single work. Also, the result is published with a limited amount of data. There are multiple reasons for the limited work; one of the key reason is the availability of the annotated data for the research Thorson et al. (2013); Ahmed, Pasquier, and Qadah (2013). Chapman et al. highlights the problem of getting labelled data for NLP related problem Chapman et al. (2011).

Researchers are dependent on in-house or crowd-based data annotation. Recently, Alam et al. uses a crowd-based annotation technique and asks people to volunteer for data annotation, but there is no significant success in getting a large number of labelled data Alam et al. (2020). The current annotation technique is dependent on the background expertise of the annotators. On the other hand, finding the past data on an incident like mob lynching, disaster is challenging because of data restrictions by social media platforms. It requires looking at massive posts, news articles with an intensive amount of manual work. Billions of social media posts are sampled to a few thousand posts for data annotation either by random sample or keyword sampling, which brings a sampling bias in the data.

With the in-house data annotation, Forbush et al. mentions that it's challenging to hire annotator with background expertise in a domain. Another issue is the development of a codebook with a proper explanation Forbush et al. (2013). The entire process is financially costly and time taking Duchenne et al. (2009). The problem with the crowd-based annotation tools like AMT is that the low cost may result in wrong labelling of data. Many annotators who cheat, not performing the job, but using robots or answering randomly Fort, Adda, and Cohen (2011); Sabou et al. (2014).

With the emergence of social media as a news resources Caumont (2013), many people or group of people use it for different purpose like news sharing, personal opinion, social crime in the form of hate speech, cyber bullying. Nowadays, the journalists cover some of the common issues like misinformation, mob lynching, hate speech, and they also link the social media post in the news articles Cui and Liu (2017). In the proposed framework, We used the news articles from profession news website for developing the proposed framework. We only collect the news articles/blog from the credible source which does not compromise with the news quality Meyer (1988). In the next section, we discuss the proposed methodology for the AMUSED framework.

Data on Social Media Platforms

Social Media platform allows users to create and view posts in multiple formats. Every day billions of posts containing images, text, videos are shared on social media sites such as Facebook, Twitter, YouTube or Instagram Aggarwal (2011). People use a combination of image, text and video for more creative and expressive forms of communication. Data are available in different formats and each social media platform apply restriction on data crawling. For instance, Facebook allows crawling data only related to only public posts and groups. Giglietto, Rossi, and Bennato discuss the requirement of multi-modal data for the study of social phenomenon Giglietto, Rossi, and Bennato (2012). In the following paragraph, We highlighted the data format and restriction on several social media platforms.

Text Almost every social media platform allows user to create or respond to the social media post in text. But each social media platform has a different restriction on the size of the text. Twitter has a limit of 280 characters, while on YouTube, users are allowed to comment up to a limit of 10000 characters. Reddit allows 40,000 characters; Facebook has a limit of 63206 Characters, Wikipedia has no limit and so on. The content and the writing style changes with the character limit of different social media platform.

Image Like text, image is also a standard format of data sharing across different social media platforms. These platforms also have some restriction on the size of the image. Like Twitter has a limit of 5 Megabytes, Facebook and Instagram have a limit of 30 Megabytes, Reddit has a limit of 20 Megabytes. Images are commonly used across different platform. It is common in Social media platforms like Instagram, Pinterest.

Video Some platforms are primarily focused on video like YouTube. While other platforms are multi-modal which allows video, text and image. For video also there are restrictions in terms of duration like YouTube has a capacity of 12 hours, Twitter allows 140 seconds, Instagram has a limit of 120 seconds, and Facebook allows videos up to 240 minutes. The restriction of video's duration on different platforms catches different users. For instance, on Twitter and Instagram users post video with shorter duration. In contrast, YouTube has users from media organisation, vlog writer, educational institution etc where the duration of video is longer.

Problems of Current Annotation Technique

In the current annotation scenario, researchers collect the data from social media platforms for a particular issue with different search criteria. There are several problems with the current annotation approaches; some of them are highlighted below.

- First, social media platforms restrict users to fetch old data; for example, Twitter allows us to gather data only from the past seven days using the standard APIs. We need to start on-time crawling; otherwise, we lose a reasonable amount of data which also contains valuable content.
- Second, if the volume of data is high, it requires filtering based on several criteria like keyword, date, location etc. These filtering further degrades the data quality by excluding the major portion of data. For example, for hate speech, if we sample the data using hateful keyword, then we might lose many tweets which are hate speech but do not contain any hateful word.
- Third, getting a good annotator is a difficult task. Annotation quality depends on the background expertise of the person. Even we hire annotator in our organisation; we have to train them for using the test data. For crowdsourcing, maintaining annotation quality is more complicated. Moreover, maintaining a good agreement between multiple annotators is also a tedious job.
- Fourth problem is the development of annotation guidelines. We have to build descriptive guidelines for data annotation, which handle a different kind of contradiction. Writing a good codebook requires domain knowledge and consultant from experts.
- Fifth, overall, data annotation is a financially costly process and time-consuming. Sorokin and Forsyth highlighted the issue of cost while using a crowd-based annotation technique Sorokin and Forsyth (2008).
- Sixth, social media is available in multiple languages, but much research is limited to English. Data annotation in other languages, especially under-resourced languages is difficult due to the lack of experienced annotators. The difficulty adversely affects the data quality and brings some bias in the data.

In this work, we propose a framework to solve the above problems by crawling the embedded social media posts from the news articles and a detailed description is given in the proposed method section.

Proposed Method

In this section, we discuss the proposed methodology of the annotation framework. Our method consists of nine steps, they are discussed below-

Step 1: Domain Identification The first step is the identification of the domain in which we want to gather the data. A domain could focus on a particular public discourse. For example, a domain could be fake news in the US election, hate speech in trending hashtags on Twitter like #BlackLivesMatter, #riotsinsweden etc. Domain selection helps to focus on the relevant data sources.

Step 2: Data source After domain identification, the next step is the identification of data sources. Data sources may consist of either the professional news websites or the blogs that talk about the particular topic, or both. For example, many professional websites have a separate section which discusses the election or other ongoing issues. In the step, we collect the news website or blog which discuss the chosen domain.

Step 3: Web scraping In the next step, we crawl all news articles from a professional news website or blogs which discuss the domain from each data source. For instances, a data source could be Snopes Snopes (2020) or Poynter Institute (2020). We fetch all the necessary details like the published date, author, location, news content.

Step 4: Language Identification After getting the details from the news articles, we check the language of the news articles. We use ISO 639-1 codes Wikipedia (2020) for naming the language. Based on the language, we can further filter the group of news articles based on spoken language from a country and apply a language-specific model for finding meaning insights.

Step 5: Social Media Link From the crawled data, we fetch the anchor tag($\langle a \rangle$) mentioned in the news content, then we filter the hyperlinks to identify social media platforms like Twitter and YouTube. From the filtered link, we fetch unique identifiers to the social media posts, for instance, for a hyperlink consisting of tweet ID, we fetch the tweet id from the hyperlink. Similarly, we fetch the unique ID to social media for each platform. We also remove the links which are not fulfilling the filtering criteria.

Step 6: Social Media Data Crawling In this step, we fetch the data from the respective social media platform. We build a crawler for each social media platform and crawl the details using unique identifiers or Uniform Resource Locator (URL) obtained from the previous step. Due to the data restriction, we use Crowdtangle Team (2020) to fetch them from Facebook and Instagram posts. Example- for Twitter, we use Twitter crawler using tweet ID (unique identifier), we crawl details about the tweets.

Step 7: Data Labelling In this step, we assign labels to the social media data based on the label assigned to the news articles by journalists. Often news articles describe the social media post to be hate speech, fake news, or propaganda. We assign the class of the social media post mentioned in the news article as a class described by the journalist. For example, if a news article **A** containing social media post **S** has been published by a journalist **J** and journalist **J** has described the social media post **S** to be a Fake News, We label the social media post **S** as Fake News. Usually, the news article is published by a domain expert, and it assures that social media post embedded or linked in the news article is correctly labelled.

Step 8: Human verification In the next step, to check the correctness, a human verifies the assigned label to the social media post and with label mentioned in the news articles. If the label is wrongly assigned, then data is removed from the corpus. This step assures that the collected social media post contains the relevant post and correctly given label. A human can verify the label of the randomly selected news articles.

Step 9: Data Enrichment In this, we merge the social media data with the details from the news articles. It helps to accumulate extra information which might allow for further analysis. Data merging provides analysis from news authors and also explains label assigned to the social media post.

Application Domain

In this section, we consider the possible application domains of the proposed framework. Nevertheless, the framework is a general one, and it can be tailored to suit varied unmentioned domains as well where the professional news website or blogs covers the incident like election, fake news etc.

Misinformation

”Fake News is an information that is intentionally, and verifiable false and could mislead readers”Allcott and Gentzkow (2017). Misinformation is part of fake news which is created deliberately intended to deceive. There is an increasing amount of misinformation in the media, social media, and other web sources. In recent years, much research has been done for fake news detection and debunking of fake news Zhou and Zafarani (2018). In the last two decades, there is a significant increase in the spread of misinformation. Nowadays more than 100 fact-checking websites are working to tackle the problem misinformation Cherubini and Graves (2016). Fact-checking websites can help to investigate claims and assist citizens in determining whether the information used in an article is true or not.

In a real-world scenario, people spread a vast amount of misinformation during the time of a pandemic, an election or a disaster. Gupta et al. (2013). There is a 3V problem of fake news – volume – a large number of fake news, Velocity – during the peak the speed of propagation also intensifies, Variety – different formats of data like images, text, videos are used in fake news. Still, fake news detection requires a considerable effort to verify the claims.

One of the most effective strategies for tackling this problem is to use computational methods to detect false news. Misinformation has attracted significant attention in recent years as evidenced in recent publications Li et al. (2012); Li, Meng, and Yu (2011); Li et al. (2016); Popat et al. (2016). Additionally, Misinformation is adopted across language borders and consequently often spread around the globe. For example- One fake news ”Russia released lions to implement the lockdown during COVID-19” was publicised across multiple countries in different languages like Italian and Tamil Poynter (2020).

Mob Lynching

Mob lynching is a violent human behaviour where a group of people execute the legal practice without a trial which ends with a significant injury or death of a person Apel (2004). It is a worldwide problem, the first case executed in the 15th Century in Ireland, then it was trending in the USA during the 18-19th century. Often, mob lynching is initiated by rumours or fake news which gets triggered by the social media by a group of peopleArun (2019). The preventive measures taken by the government to overcome all obstacles and prevent further deaths were not successful in its entirety.

Getting the data for analysis of mob lynching is difficult because of the unexpected events occurring throughout the year, mainly in remote areas. There is no common search term or keyword that helps to crawl social media. So, if we fetch the specific social media post from the news articles which is covering analysis about the mob lynching Arun (2019), we can use it for several studies. It will also help to analyse the cause and pattern from the previous incident Griffin (1993).

Online Abuse

Online abuse is any kind of harassment, racism, personal attacks, and other types of abuse on online social media platforms. The psychological effects of online abuse on individuals can be extreme and lasting Mishra, Yannakoudakis, and Shutova (2019). Online abuse in the form of hate speech, cyberbullying, personal attacks are common issue Mishra, Yannakoudakis, and Shutova (2019). Many research has been done in English and other widely spoken languages, but under-resourced languages like Hindi, Tamil (and many more) are not well explored. Gathering data in these languages is still a big challenge, so our annotation framework can easily be applied to collect the data on online abuse in multiple languages.

Implementation

In this, we discuss the implementation of our proposed framework. As a case study, we apply the AMUSED for data annotation for COVID-19 misinformation in the following way:

Step 1: Domain Identification Out of several possible application domains, we consider the spread of misinformation in the context of COVID-19. We choose this the topic since because, December 2019, the first official report of COVID-19, misinformation spreading over the web Shahi and Nandini (2020). The increase of misinformation is one of the big problems during the COVID-19 problems. The director of the World Health Organization(WHO), considers that with COVID, we are fighting with both Pandemic and Infodemic The Guardian (2020). Infodemic is a word coined by World Health Organization (WHO) to describe the misinformation of virus, and it makes hard for users to find trustworthy sources for any claim made on the COVID-19 pandemic, either on the news or social media World Health Organization and others (2020); Zarocostas (2020).

One of the fundamental problems is the lack of sufficient corpus related to pandemic Shahi, Dirkson, and Majchrzak (2020). Content of the misinformation depends on the domain; for example, during the election, we have a different set of misinformation compared to a pandemic like COVID-19, so domain identification helps to focus on specif topic.

Step 2: Data Sources For data source, we looked for 25 fact-checking websites(like Politifact, Boomlive) and decided to use the Poynter and Snopes. We choose Poynter

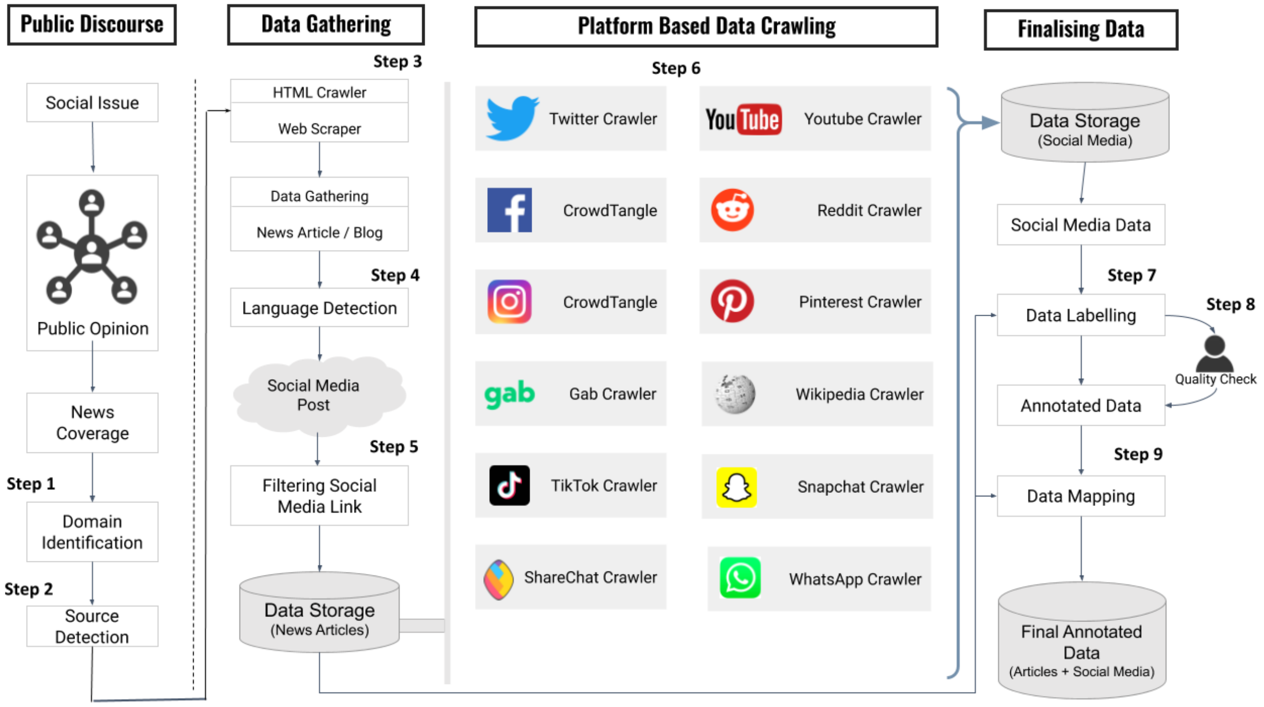


Figure 1: AMUSED: An Annotation Framework for Multi-modal Social Media data

because Poynter has a central data hub which collects data from more 98 fact-checking websites while Snopes is not integrated with Poynter but having more than 300 fact-checked articles on COVID-19. We describe the two data sources as follow-

Snopes- Snopes Snopes (2020) is an independent news house owned by Snopes Media Group. Snopes verifies the correctness of misinformation spread across several topics like election, COVID-19. As for the fact-checking process, they manually verify the authenticity of the news article and performs a contextual analysis. In response to the COVID-19 infodemic, Snopes provides a collection of a fact-checked news article in different categories based on the domain of the news article.

Poynter- Poynter is a non-profit making institute of journalists Institute (2020). In COVID-19 crisis, Poynter came forward to inform and educate to avoid the circulation of the fake news. Poynter maintains an International Fact-Checking Network(IFCN), the institute also started a hashtag #CoronaVirusFacts and #DatosCoronaVirus to gather the misinformation about COVID-19. Poynter maintains a database which collects fact-checked news from 98 fact-checking organisation in 40 languages.

Step 3: Web Scraping In this step, we developed a Python-based crawler using Beautiful soup Richardson (2007) to fetch all the news articles from the Poynter and Snopes. Our crawler collects important information like the title of the news articles, name of the fact-checking websites, date of publication, the text of the news articles, and a class of news articles. We have assigned a unique identifier to each of them and its denoted by FCID. A short description of each element given in table 1.

Step 4: Language Detection We collected data in multiple languages like English, German, Hindi etc. To identify the language of the news article, we have used langdetect Shuyo (2014), a Python-based library to detect the language of the news articles. We used the textual content of new articles to check the language of the news articles. Our dataset is categorise into 40 different languages.

Step 5: Social Media Link In the next step, while doing the HTML crawling, we filter the URL from the parsed tree of the DOM (Document Object Model). We analysed the URL pattern from different social media platforms and applied keyword-based filtering from all hyperlinks in the DOM. We store that URLs in a separate column as the social media link. An entire process of finding social media is shown in figure 2. Some of the URL patterns are discussed below-
Twitter- For each tweet, Twitter follows a pattern twitter.com/user_name/status/tweetid. So, in the collection hyperlink, we searched for the keyword, ”twitter.com” and ”status”, it assures that we have collected the hyperlink which referring to the tweet.
YouTube- For each YouTube video, YouTube follows a pattern hwww.youtube.com/watch?v=vidoeid. So, in the collection hyperlink, we searched for the keyword, ”youtube.com” and ”watch”, these keyword assures that we have collected the hyperlink which referring to the particular YouTube Video.
Reddit- For each subreddit, Reddit follows a pattern www.reddit.com/r/subreddit_topic/. So, in the collection hy-

Element	Definition	Example
News_ID	We provide a unique identifying ID to each news articles. We use acronym for news source and the number to identify a news articles.	PY9
Newssource_URL	It is a unique identifier pointing to the news articles.	`https://factcheck.afp.com/video-actually-shows-anti-government-protest-belarus`
News_Title	In this field, we store the title of the news articles.	A video shows a rally against coronavirus restrictions in the British capital of London.
Published_date	Each news articles published the fact check article with a class like false, true, misleading. We store it in the class column.	01 September 2020
News_Class	We provide a unique identifying ID to each news articles.	False
Published-By	In this field, we store the name of the professional news websites or blog, for example, AFP, Quint etc.	
Country	Each news articles published the fact check article with a class like false, true, misleading. We store it in the class column.	Australia
Language	We provide a unique identifying ID to each news articles.	English

Table 1: Name, definition and an example of elements collected from new articles.

perlink, we searched for the keyword, ”reddit.com” and a regex code to detect ”reddit.com/r/”, which confirms that we have collected the hyperlink which referring to the particular subreddit.

Similarly, we followed the approach for other social media platforms like Facebook, Instagram, Wikipedia, Pinterest, Gab. In the next step, we used the regex code to filter the unique ID for each social media post like tweet ID for Twitter, Video ID for YouTube.

Step 6: Social Media Data Crawling After web scraping, we have the unique identifier of each social media post like tweet ID for Twitter, Video Id for videos etc. We build a Python-based program for crawling the data from the respective social media platform. We describe some of the crawling tool and the details about the collected data.

Twitter- We used Python crawler using Tweepy Roesslein (2020), which crawls all details about a Tweet. We collect text, time, likes, retweet, user details such as name, location, follower count.

YouTube- For YouTube, we built a Python-based crawler which collects the textual details about the video, like title, channel name, date of upload, likes, dislikes. We also crawled the comments of the respective.

Similarly, we build our crawler for other platforms, but for Instagram and Facebook, we used the CrowdTangle for data crawling, data is limited to posts from public pages and group Team (2020).

Step 7: Data Labelling For data labelling, we used the label assigned in the news articles then we map the social media post with their respective news article and assign the label to the social media post. For example, a tweet extracted from news article is mapped to the class of the news article. An entire process of data annotating shown in figure 3.

Step 8: Human Verification In the next step, we manually overlook each social media post to check the correctness of the process. We provided the annotator with all necessary information about the class mapping and asked them to verify it. For example- In figure 3, human open the news article using the newssource_URL and verified the label assigned to the tweet. For COVID-19 misinformation, A human checked randomly sampled 10% social media post from each social media platforms and verified the label assign to the social media post and label mentioned in the news articles. With the random checks, we found that all the assigned labels are correct. This helps make sure the assigned label is correct and reduces the bias or wrongly assigned label. We further normalise the data label into false, partially false, true and others using the definitions mentioned in Shahi, Dirkson, and Majchrzak (2020). The number of social media post found in four different category is shown in Table 3.

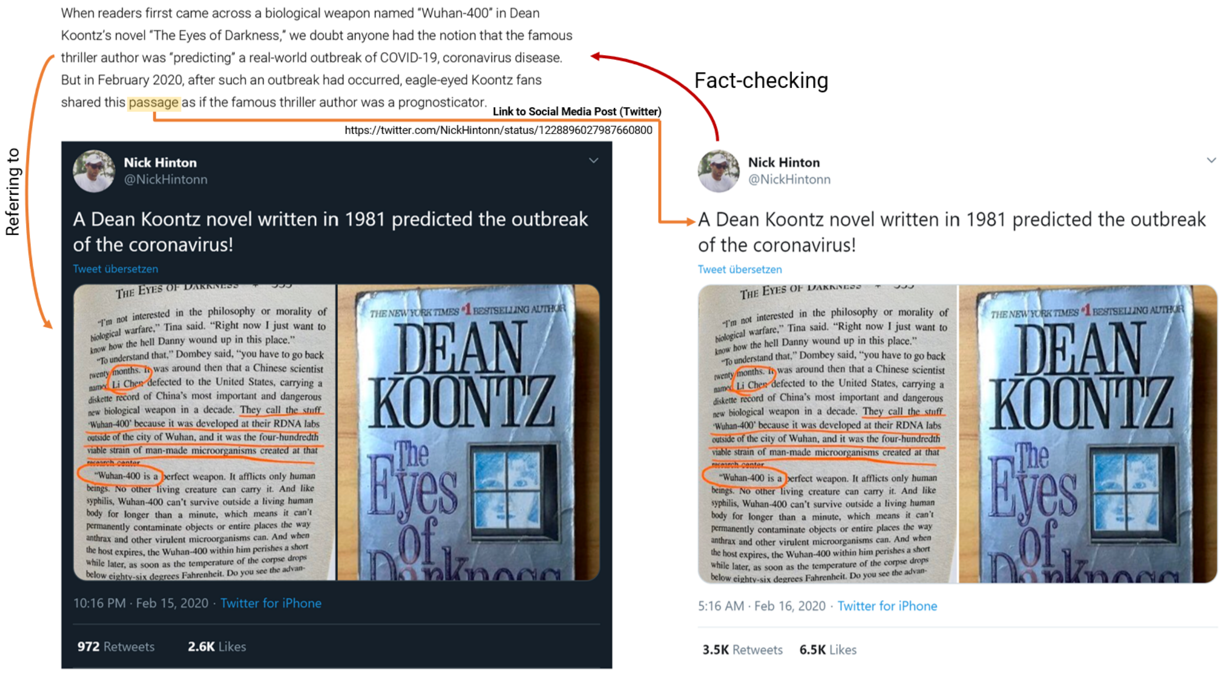
Step 9: Data Enrichment In this step, we enrich the data by providing extra information about the social media post. The first step is merging the social media post with the respective news article, and it includes additional information like textual content, news source, author.

The detailed analysis of the collected data is discussed in the result section. Based on the results, we also discuss some of the findings in the discussion section. A snapshot of the labelled data from Twitter is shown in figure 4. We will release the data as open-source for further study.

Results

For the use case of misinformation on COVID-19, we identified IFCN as the data source, and we collected data from different social media platforms. We found that around 51% of news articles contain linked their content to social media websites. Overall, we have collected 8077 fact-checked news articles from 105 countries in 40 languages. A detailed description of social media data extracted using the AMUSED framework is presented in table 2.

We have cleaned the hyperlinks collected using the AMUSED framework. We filtered the social media posts by removing the duplicates using a unique identifier of social media post. We have presented a timeline plot of data collected from different social media platforms in figure 5. We plotted the data from those social media platform which has



When readers first came across a biological weapon named "Wuhan-400" in Dean Koontz's novel "The Eyes of Darkness," we doubt anyone had the notion that the famous thriller author was "predicting" a real-world outbreak of COVID-19, coronavirus disease. But in February 2020, after such an outbreak had occurred, eagle-eyed Koontz fans shared this passage as if the famous thriller author was a prognosticator.

Referring to
Fact-checking
Link to Social Media Post (Twitter)
https://twitter.com/NickHintonn/status/1228896027987660800

Nick Hinton
@NickHintonn

A Dean Koontz novel written in 1981 predicted the outbreak of the coronavirus!

Tweet übersetzen

THE EYES OF DARKNESS

"I'm not interested in the philosophy or morality of biological warfare," Tina said. "Right now I just want to know how the hell Danny wound up in this place."

"To understand that," Dombey said, "you have to go back twenty months. It was around then that a Chinese scientist named Li Chen defected to the United States, carrying a diskette record of China's most important and dangerous new biological weapon in a decade. They call the stuff 'Wuhan-400' because it was developed at their RDNA labs outside of the city of Wuhan, and it was the four-hundredth viable strain of man-made microorganisms created at that research center.

"Wuhan-400 is a perfect weapon. It afflicts only human beings. No other living creature can carry it. And like syphilis, Wuhan-400 can't survive outside a living human body for longer than a minute, which means it can't permanently contaminate objects or entire places the way anthrax and other virulent microorganisms can. And when the host expires, the Wuhan-400 within him perishes a short while later, as soon as the temperature of the corpse drops below eighty-six degrees Fahrenheit. Do you see the advan-

10:16 PM · Feb 15, 2020 · Twitter for iPhone

972 Retweets 2.6K Likes

Nick Hinton
@NickHintonn

A Dean Koontz novel written in 1981 predicted the outbreak of the coronavirus!

Tweet übersetzen

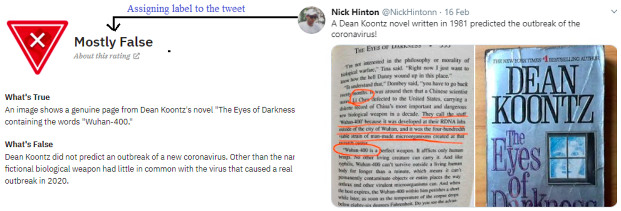
5:16 AM · Feb 16, 2020 · Twitter for iPhone

3.5K Retweets 6.5K Likes

Figure 2: An Illustration of data collection from social media platform(Twitter) Hinton (2020) from a news article Evon (2020)

SM Platform	Count of Post(Unique)	Post with Text	Post with Image	Post with Text+Image	Post with Video
Facebook	5799(3200)	1167	567	1006	460
Instagram	385(197)	-	106	41	52
Pinterest	5(3)	-	3	0	0
Reddit	67(33)	16	10	7	0
TikTok	43(18)	-	-	-	18
Twitter	3142(1758)	1300	116	143	199
Wikipedia	393(176)	106	34	20	16
YouTube	2087(916)	-	-	-	916

Table 2: Summary of data collected from different social media platforms, the number in round braces indicate the count of unique social media posts.



What's True

An image shows a genuine page from Dean Koontz's novel "The Eyes of Darkness containing the words "Wuhan-400."

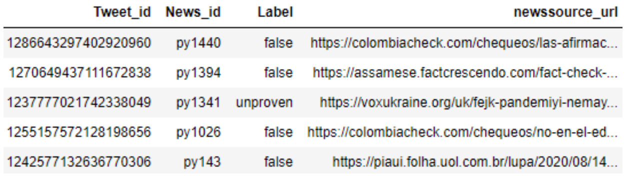
What's False

Dean Koontz did not predict an outbreak of a new coronavirus. Other than the nar fictional biological weapon had little in common with the virus that caused a real outbreak in 2020.

Nick Hinton @NickHintonn · 16 Feb

A Dean Koontz novel written in 1981 predicted the outbreak of the coronavirus!

Figure 3: An Illustration for annotation of social media posting using the label mentioned in the news article.



Tweet_id	News_id	Label	newssource_url
1286643297402920960	py1440	false	https://colombiacheck.com/chequeos/las-afirmac...
1270649437111672838	py1394	false	https://assamese.factcrescendo.com/fact-check-...
1237777021742338049	py1341	unproven	https://voxukraine.org/uk/fejk-pandemiyi-nemay...
1255157572128198656	py1026	false	https://colombiacheck.com/chequeos/no-en-el-ed...
1242577132636770306	py143	false	https://piaui.folha.uol.com.br/lupa/2020/08/14...

Figure 4: A Glimpse of annotated data collected from Twitter.

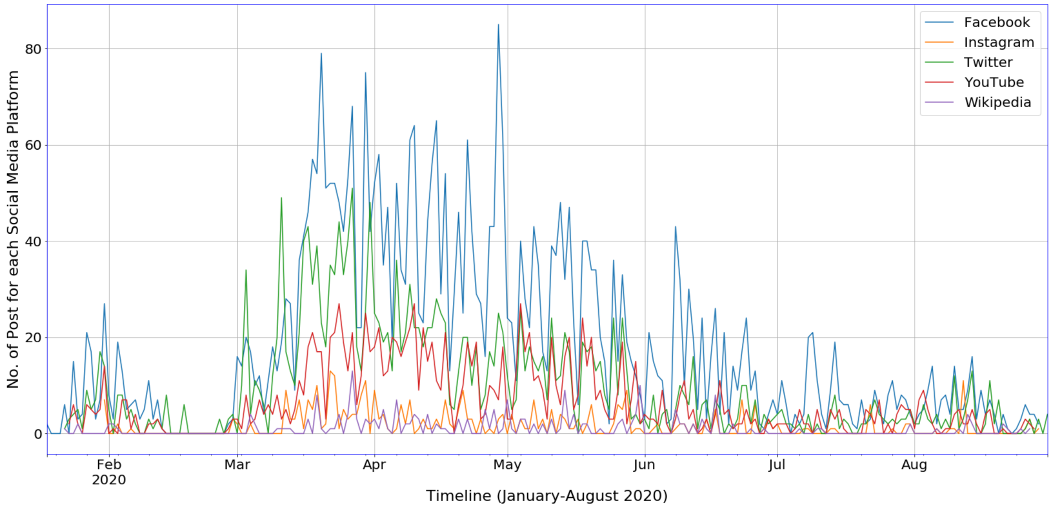


Figure 5: A timeline distribution of data collected from a number of different Social Media Platform from January 2020 to August 2020, we have presented the platform having data count more than 25.

SM Platform	False	Partially False	Other	True
Facebook	2776	325	94	6
Instagram	166	28	2	1
Pinterest	3	0	0	0
Reddit	21	9	2	1
TikTok	9	0	0	0
Twitter	1318	234	50	13
Wikipedia	154	18	3	1
YouTube	739	164	13	0

Table 3: Summary of COVID-19 misinformation data collected from different social media platforms, deleted and duplicate posts are excluded in the count.

the total number of post more than 25 unique posts in Table 3 because it depreciates the plot distribution. We dropped the plot from Pinterest(3), Whatsapp(23), Tiktok(25), Reddit(43). The plot shows that most of the social media posts are from Facebook and Twitter, then followed by YouTube, then Wikipedia and Instagram. We have also presented the class distribution of these social media post in table 3. The figure 5 shows that the number of post overall social media post was maximum during the mid-March to mid-May, 2020. Misinformation also follows the trend of the COVID-19 situation in many countries because the number of social media post also decreased after June 2020. The possible reason could be either the spread of misinformation is reduced, or fact-checking websites are not focusing on this issue as during the early stage.

Discussion

From our study, we highlighted some of the useful points. Usually, the fact-checking website links the social media post from multiple social media platforms. We tried to gather data from various social media platforms, but we found the maximum number of links from Facebook, Twitter, and YouTube. There are few unique posts from Reddit(21), TikTok(9) but they were less than what we were expecting Brennen et al. (2020). Surprisingly there are only three unique posts from Pinterest, and there are no data available from Gab, ShareChat, and Snapchat. However, Gab is well known for harmful content, and people in their regional languages use ShareChat. There are only three unique posts from Pinterest. Many people use Wikipedia as a reliable source of information, but there are 393 links from Wikipedia. Hence, overall fact-checking website is limited to some trending social media platforms like Twitter or Facebook while social media platforms like Gab, TikTok is famously famous for malformation, misinformation Brennen et al. (2020). WhatsApp is an instant messaging app, used among friends or group of people. So, we only found some hyperlink which links to the public WhatsApp group. To increase the visibility of fact-checked articles, a journalist can also use schema.org vocabulary along with the Microdata, RDFa, or JSON-LD formats to add details about misinformation to the news articles Shahi, Nandini, and Kumari (2019).

Another aspect is the diversity of social media post on the different social media platforms. More often, news articles mention Facebook, Twiter, YouTube but less number of post from Instagram, Pinterest, no post from Gab, Tiktok. There might be these platforms actively ask or involve the fact-checking website for monitoring the content on their platform, or the journalists are more focused on these platforms only. But it would be interesting to study the proposition of fake news on different platforms like Tiktok, Gab. We have also analysed the multi-modality of the data on the social media platform. In the case of misinformation on COVID-19, the amount of misinformation on text is more compare to video or image. But, in table 3 we show that apart from text, the fake news is also shared as image, video or mixed-format like image+text. It will also be beneficial to detect fake news on different platforms. It also raises the open question of cross-platform study on a particular topic like misinformation on COVID-19. Someone can also build a classification model Shahi et al. (2018); Nandini et al. (2018) to detect a class of fake news into true, false, partially false or other categories of news articles.

While applying AMUSED framework on the misinformation on COVID-19, we found that misinformation across multiple source platform, but it mainly circulated across Facebook, Twitter, YouTube. Our finding raises the concern of mitigating the misinformation on these platforms.

Conclusion and Future Work

In this paper, we presented a semi-automatic framework for social media data annotation. The framework can be applied to several domains like misinformation, mob lynching, and online abuse. As a part of the framework, we also used a python based crawler for different social media websites. After data labelling, the labels are cross-checked by a human which ensures a two-step verification of data annotation for the social media posts. We also enrich the social media post by mapping it to the news article to gather more analysis about it. The data enrichment will be able to provide additional information for the social media post. We have implemented the proposed framework for collecting the misinformation post related to the COVID-19

As future work, the framework can be extended for getting the annotated data on other topics like hate speech, mob lynching etc. AMUSED will decrease the labour cost and time for the data annotation process. AMUSED will also increase the quality of the data annotation because we crawl the data from news articles which are published by an expert journalist.

References

Aggarwal, C. C. 2011. An introduction to social network data analytics. In *Social network data analytics*. Springer. 1–15.

Ahmed, S.; Pasquier, M.; and Qadah, G. 2013. Key issues in conducting sentiment analysis on arabic social media text. In *2013 9th International Conference on Innovations in Information Technology (IIT)*, 72–77. IEEE.

Alam, F.; Dalvi, F.; Shaar, S.; Durrani, N.; Mubarak, H.; Nikolov, A.; Martino, G. D. S.; Abdelali, A.; Sajjad, H.; Darwish, K.; et al. 2020. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms. *arXiv preprint arXiv:2007.07996*.

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31(2):211–36.

Apel, D. 2004. *Imagery of lynching: Black men, white women, and the mob*. Rutgers University Press.

Aroyo, L., and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36(1):15–24.

Arun, C. 2019. On whatsapp, rumours, and lynchings. *Economic & Political Weekly* 54(6):30–35.

Braun, J., and Gillespie, T. 2011. Hosting the public discourse, hosting the public: When online news and social media converge. *Journalism Practice* 5(4):383–398.

Brennen, J. S.; Simon, F.; Howard, P. N.; and Nielsen, R. K. 2020. Types, sources, and claims of covid-19 misinformation. *Reuters Institute* 7:3–1.

Carlson, M. 2016. Embedded links, embedded meanings: Social media commentary and news sharing as mundane media criticism. *Journalism Studies* 17(7):915–924.

Caumont, A. 2013. 12 trends shaping digital news. *Pew Research Center* 16.

Chapman, W. W.; Nadkarni, P. M.; Hirschman, L.; D'avolio, L. W.; Savova, G. K.; and Uzuner, O. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.

Cherubini, F., and Graves, L. 2016. The rise of fact-checking sites in europe. *Reuters Institute for the Study of Journalism, University of Oxford. http://reutersinsfitute. polifics. ox. ac. uk/our-research/rise-fact-checking-sites-europe*.

Cook, P., and Stevenson, S. 2010. Automatically identifying changes in the semantic orientation of words. In *LREC*.

Cui, X., and Liu, Y. 2017. How does online news curate linked sources? a content analysis of three online news media. *Journalism* 18(7):852–870.

Duchenne, O.; Laptev, I.; Sivic, J.; Bach, F.; and Ponce, J. 2009. Automatic annotation of human actions in video. In *2009 IEEE 12th International Conference on Computer Vision*, 1491–1498. IEEE.

Evon, D. 2020. *Was Coronavirus Predicted in a 1981 Dean Koontz Novel?*

Forbush, T. B.; Shen, S.; South, B. R.; and DuValla, S. L. 2013. What a catch! traits that define good annotators. *Studies in health technology and informatics* 192:1213–1213.

Fort, K.; Adda, G.; and Cohen, K. B. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics* 37(2):413–420.

Giglietto, F.; Rossi, L.; and Bennato, D. 2012. The open laboratory: Limits and possibilities of using facebook, twitter, and youtube as a research data source. *Journal of technology in human services* 30(3-4):145–159.

Grant-Muller, S. M.; Gal-Tzur, A.; Minkov, E.; Nocera, S.; Kuflik, T.; and Shoor, I. 2014. Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems* 9(4):407–417.

Griffin, L. J. 1993. Narrative, event-structure analysis, and causal interpretation in historical sociology. *American journal of Sociology* 98(5):1094–1133.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, 729–736.

Hinton, N. 2020. *A Dean Koontz novel [...].*

Institute, P. 2020. *The International Fact-Checking Network.*

Li, X.; Dong, X. L.; Lyons, K.; Meng, W.; and Srivastava, D. 2012. Truth finding on the deep web: Is the problem solved? *Proceedings of the VLDB Endowment* 6(2).

Li, Y.; Gao, J.; Meng, C.; Li, Q.; Su, L.; Zhao, B.; Fan, W.; and Han, J. 2016. A survey on truth discovery. *ACM Sigkdd Explorations Newsletter* 17(2):1–16.

Li, X.; Meng, W.; and Yu, C. 2011. T-verifier: Verifying truthfulness of fact statements. In *2011 IEEE 27th International Conference on Data Engineering*, 63–74. IEEE.

Meyer, P. 1988. Defining and measuring credibility of newspapers: Developing an index. *Journalism quarterly* 65(3):567–574.

Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Nandini, D.; Capecci, E.; Koefoed, L.; Laña, I.; Shahi, G. K.; and Kasabov, N. 2018. Modelling and analysis of temporal gene expression data using spiking neural networks. In *International Conference on Neural Information Processing*, 571–581. Springer.

Popat, K.; Mukherjee, S.; Strötgen, J.; and Weikum, G. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2173–2178.

Poynter. 2020. Russia released 500 lions.

Richardson, L. 2007. Beautiful soup documentation. *April*.

Roesslein, J. 2020. http://www.tweepy.org.

Sabou, M.; Bontcheva, K.; Derczynski, L.; and Scharl, A. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, 859–866.

Shahi, G. K., and Nandini, D. 2020. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*.

Shahi, G. K.; Bilbao, I.; Capecci, E.; Nandini, D.; Choukri, M.; and Kasabov, N. 2018. Analysis, classification and marker discovery of gene expression data with evolving spiking neural networks. In *International Conference on Neural Information Processing*, 517–527. Springer.

Shahi, G. K.; Dirkson, A.; and Majchrzak, T. A. 2020. An exploratory study of covid-19 misinformation on twitter. *arXiv preprint arXiv:2005.05710*.

Shahi, G. K.; Nandini, D.; and Kumari, S. 2019. Inducing schema. org markup from natural language context. In *Workshop Papers of LuxLogAI*, volume 10, 38–42.

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

Shuyo, N. 2014. Language-detection library.

Snopes. 2020. Collections archive.

Sorokin, A., and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, 1–8. IEEE.

Stieglitz, S.; Mirbabaie, M.; Ross, B.; and Neuberger, C. 2018. Social media analytics–challenges in topic discovery, data collection, and data preparation. *International journal of information management* 39:156–168.

Talwar, S.; Dhir, A.; Kaur, P.; Zafar, N.; and Alrasheedy, M. 2019. Why do people share fake news? associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services* 51:72–82.

Team, C. 2020. Crowdtangle. facebook, menlo park, california, united states.

The Guardian. 2020. The WHO v coronavirus: why it can't handle the pandemic.

Thorson, K.; Driscoll, K.; Ekdale, B.; Edgerly, S.; Thompson, L. G.; Schrock, A.; Swartz, L.; Vraga, E. K.; and Wells, C. 2013. Youtube, twitter and the occupy movement: Connecting content and circulation practices. *Information, Communication & Society* 16(3):421–451.

Wikipedia. 2020. *List of ISO 639-1 codes.*

World Health Organization and others. 2020. Coronavirus disease 2019 (covid-19): situation report, 72. Technical report, World Health Organization.

Zarocostas, J. 2020. World Report How to fight an infodemic. *The Lancet* 395:676.

Zhou, X., and Zafarani, R. 2018. Fake news: A survey of research, detection methods, and opportunities. *CoRR* abs / 1812.00315.