Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection

William Scott Paka [a,*,1], Rachit Bansal [b,1,2], Abhay Kaushik [c,2], Shubhashis Sengupta [d], Tanmoy Chakraborty [a]

[a] IIIT-Delhi, India
[b] DTU-Delhi, India
[c] IIT-Kanpur, India
[d] Accenture Labs, India

ABSTRACT

As the COVID-19 pandemic sweeps across the world, it has been accompanied by a tsunami of fake news and misinformation on social media. At the time when reliable information is vital for public health and safety, COVID-19 related fake news has been spreading even faster than the facts. During times such as the COVID-19 pandemic, fake news can not only cause intellectual confusion but can also place people's lives at risk. This calls for an immediate need to contain the spread of such misinformation on social media. We introduce CTF, a large-scale COVID-19 Twitter dataset with labelled genuine and fake tweets. Additionally, we propose **Cross-SEAN**, a cross-stitch based semi-supervised end-to-end neural attention model which leverages the large amount of unlabelled data. Cross-SEAN partially generalises to emerging fake news as it learns from relevant external knowledge. We compare Cross-SEAN with seven state-of-the-art fake news detection methods. We observe that it achieves 0.95 F1 Score on CTF, outperforming the best baseline by 9%. We also develop **Chrome-SEAN**, a Cross-SEAN based chrome extension for real-time detection of fake tweets.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The increase in accessibility to Internet has dramatically changed the way we communicate and share ideas. Social media consumption is one of the most popular activities online. Nowadays, it is a trend to rely on such platforms for news updates. The absence of a verification barrier allows misinformation on sites online. Due to the complexity of the issue, the definition of "fake news" is not well defined. A few definitions used in prior studies are as follows: 'A news article that is intentionally and verifiably false' [1,2] relating to news that are deceptive in nature, 'A news article or message published and propagated through media, carrying false information regardless of the means and motives behind it' relating to various forms of false news and misinformation [3–7]. A few broader definitions by Zhou et al. [8] state, 'Fake news is false news', 'Fake news is intentionally false news

published by a news outlet.' For our purpose, we define COVID-19 fake tweet as any tweet with information which contradicts the statements released by the governmental health organisations,[3] and genuine tweets to be the tweets obtained from their official accounts.

On 30 January 2020, The World Health Organisation (WHO) has declared COVID-19 to be a Public Health Emergency of International Concern and issued a set of Temporary Recommendations. A recent study observed 25% increase in average user social media activity due to the global lockdown [9]. UNESCO stated, "during this coronavirus pandemic, fake news is putting lives at risk". Fake news, ranging from the speculations around origin of the virus to baseless prevention and cures, is spreading rapidly without any valid evidence. WHO has recently declared the spread of COVID-19 related misinformation as an 'Infodemic'; according to their definition, "An infodemic is an overabundance of information, both online and offline. It includes deliberate attempts to disseminate wrong information to undermine the public health response and advance alternative agendas of groups or individuals". WHO, CDC (Centers for Disease Control and Prevention) and other government bodies have set up specific web pages in order to curb major misconceptions about the virus

---

* Corresponding author.
E-mail addresses: william18026@iiitd.ac.in (W.S. Paka), rachitbansal_2k18ee152@dtu.ac.in (R. Bansal), kabhay@iitk.ac.in (A. Kaushik), shubhashis.sengupta@accenture.com (S. Sengupta), tanmoy@iiitd.ac.in (T. Chakraborty).

[1] Equal contribution.
[2] Work done when the authors were interns at IIIT-Delhi.

[3] https://en.wikipedia.org/wiki/List_of_health_departments_and_ministries.
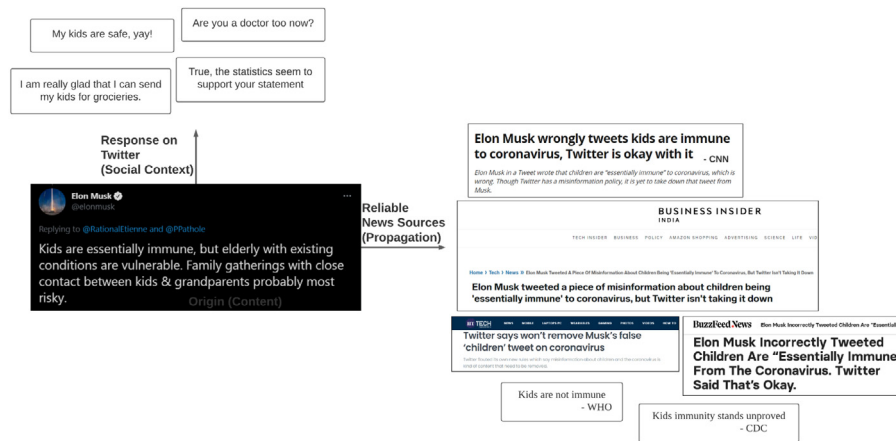
Fig. 1. An example of *origin*, *propagation* and *social context* of a popular misinformation. The responses for a tweet with misinformation seem to be coherent to it, and could ultimately spread it wider and deeper into the follower networks. Both the tweets and responses contradict the reliable news source.

and to maintain public awareness. Any single false news that gains enormous traction can negate the significance of a body of verified facts. When a tweet with misinformation is retweeted by an influential person or by a verified account, the marginal impact grows largely. The analysis, identification, and elimination of fake news thus have become a task of utmost importance. Therefore, there is an immediate need to detect the fake news and stop their spreading.

Till now, no verification barrier exists that can authenticate the content being shared on social media platforms. Due to this, quite often, general people are misinformed when an unreliable news or information is shared irrespective of intentions. With increase in reliance on social media platforms such as Twitter and Facebook for information, the spread of misinformation also tends to increase. Fake news is usually targeted for financial or political gain with click-bait titles or advertisement links gaining user attention. The spread of fake news is proven to be a threat in the past during global events such as US 2016 elections and the Brexit. Studies showed that automated bots are used for spreading fake content [10]; however, all the posts of bots cannot be considered as fake since they are devised to post non-fake content too. Genuine users seldom fall prey to fake content, and with uninformed knowledge sharing among their network makes genuine users major contributors to its spread.

Twitter is one of the largest micro-blogging platforms with over 1.5M daily active users combating fake news since a long time. The major exploitation of fake news is highlighted during the 2016 U.S. presidential election campaign. The existence of 'echo chamber effect' on social media allows biased information to be spread wider and deeper [11]. Tweets containing fake content show far wider reach, spreading rapidly than normal tweets, and such variations in propagation can be clearly observed in tweets related to political news. Such tweet propagation behaviour is partly due to the innate nature of users to retweet content which is provocative, aligning to their beliefs, irrespective of the truthfulness of the content. Social and psychological factors with 'valence effect' [12] play an important role in the spread of fake news. Studies also showed the involvement of bots to create and spread fake news [10]. News involving any political figure in power create huge fluctuations in stock markets and trades economically. For example, a 2013 tweet 'Breaking: Two Explosions in the White House and Barack Obama is injured', from a hacked Associated Press account created a loss of $136 billion worth of stock value [13]. Twitter has a long history of accounts getting stolen, and hackers with motivations to create mass hysteria take control of verified accounts for wider spread

of hoax. Although the character limit helps the amount of textual content being shared, other forms of content such as images, videos and links are also exploited to spread false information. Twitter usually deletes tweets and users that are flagged post-verification; however, this is not a scalable solution for automated fake news verification.

Due to the lockdown and work from home conditions during COVID-19 pandemic, Twitter witnessed a 30% rise in daily average usage. With isolation from the external world, users turn to social media platforms for any updates related to the pandemic. Due to uninformed knowledge, users tend to retweet content which may not be totally accurate. At the beginning of the pandemic, very limited information is available to the public on the realities of the virus. Even verified users such as Elon Musk tweeted stating that "Kids are essentially immune" which provides statistical evidence in which there are no infected people below the age 19. Public health experts later released a statement debunking his claim. We illustrate this in Fig. 1, showing the tweet with misinformation by Elon Musk along with the 'responses received in Twitter' and 'reliable news sources statements'. We can notice that the retweets are coherent to the misinformed tweet which spread the misinformation across other networks, wider and deeper. The news from verified sources state otherwise, clearly debunking the said statement Due to the scarcity of reliable information source, multiple fact checking sites depend on statements released by Public Health bodies. Although few users tweet and retweet false content without any ill-intention, there exist users who create and spread false news for political gains. Diffusion of fake tweets and genuine tweets vary in a pandemic setting such as this [14]. Tweeting a political tweet with false information multiple times from several accounts with various trending hashtags, called 'Hashtag hijacking' is also observed. Fig. 2(a) shows the count of favourites and retweets for both genuine and fake tweets, whereas Fig. 2(b) shows the friends and followers count of users posting genuine and fake tweets. We can clearly observe from Fig. 2(a) that genuine tweets tend to have higher favourite count compared to retweet count whereas the fake tweets tend to have higher retweet count, propagating the false information to a wider range. We can also observe from Fig. 2(b) that users posting genuine content have higher number of friends than followers, and users posting fake content have higher number of followers than friends – this setting again allows spread of fake news towards larger audiences through the users posting fake content.

The rest of our paper is organised as follows. We discuss related works on fake news detection and semi-supervised models for text classification in Section 2. Section 3 describes our
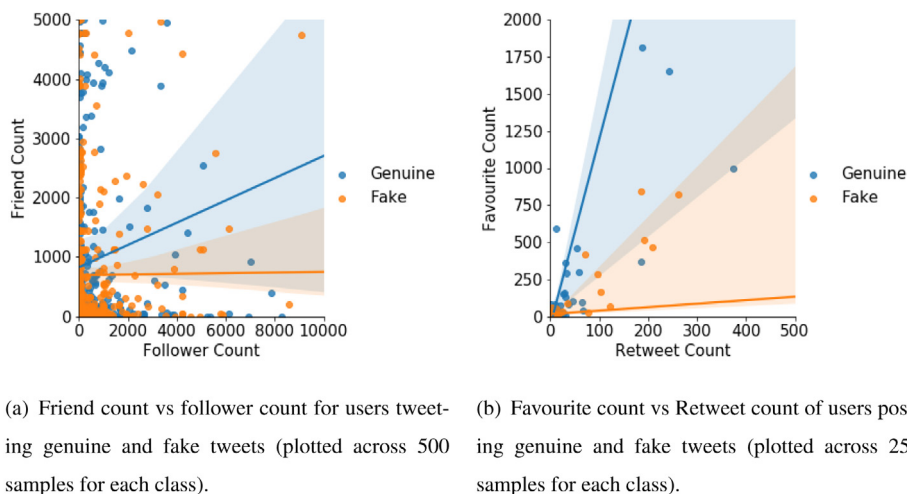
(a) Friend count vs follower count for users tweeting genuine and fake tweets (plotted across 500 samples for each class).

(b) Favourite count vs Retweet count of users posting genuine and fake tweets (plotted across 250 samples for each class).

Fig. 2. Correlations between (a) user features and (b) tweet features for genuine and fake tweets. In (b), note that a large number of samples are present close to the origin.

four-stage dataset collection and annotation processes, which is followed by further analysis of the dataset on various aspects in Section 4. The proposed Cross-SEAN model and training strategies are introduced in Section 5, while its evaluation and a detailed ablation study are shown in Section 6. For real-time usage of Cross-SEAN, the developed chrome extension, Chrome-SEAN and the user study are described in Section 7. Finally, the paper is concluded with discussions, shortcomings and future work in Section 8.

Our contribution I: CTF- A COVID-19 fake news dataset and its analysis. With the aforementioned concerns, it is evident that more research is required to detect and neutralise fake tweets and keep the users warned. Although research communities are interested to work on the challenging task of COVID-19 fake news detection which is one of the pressing issues of our time, the absence of a publicly available labelled COVID-19 misinformation dataset is a major bottleneck to design automated detection models. Also, not everyone possesses the resources to collect such a dataset, as it is cumbersome. We fill this gap by introducing CTF, the first COVID-19 Twitter fake news dataset, consisting of a mixture of both labelled and unlabelled tweets. Our dataset contains a total of $45.26K$ labelled tweets, among which $18.55K$ are labelled as 'genuine' and $26.71K$ as 'fake'. In addition, it contains $21.85M$ unlabelled tweets, which can be used to enrich the diversity of the dataset, in terms of linguistic and contextual features in general. A detailed analysis of the dataset unfolds many interesting observations. E.g., fake news content tends to – (i) accompany less URLs and more multimedia content, (ii) receive much lesser likes and retweets, (iii) exhibit mostly neutral and negative sentiment, as compared to genuine content. Our dataset collection is a four stage process, starting from hydration of Tweets, collection of supporting statements, usage of fine-trained Transformer models such as BERT and RoBERTa, to manual annotation. As COVID-19 is an emerging topic, we rely on certain government health organisations and fact checking sites such as PolitiFact, Snopes, TruthOrFiction, etc, which release statements on widely popular misconceptions. We then use tweets on the collected facts using BERT and RoBERTa to identify supporting or contradicting claims, which are then partially annotated. The major part of our genuine tweets are taken from governmental health organisations.

Our contribution II: Cross-SEAN. Two major issues in any fake news detection task are the lack of labelled data to train a deep neural model and the inability to detect fake news that are different from the training data (emerging fake news). To address these issues, we propose Cross-SEAN, a cross-stitch based semi-supervised attention neural model. Cross-SEAN works in a semi-supervised way leveraging the vast unlabelled data to learn the writing style of tweets in general. It considers user metadata, tweet metadata, and external knowledge in addition to tweet text as its inputs. External knowledge is collected on the fly in the form of stances close to tweets from trusted domains and allows a way for Cross-SEAN to not restrict to the train data, as external knowledge can contain information which is absent in the train data partially helping with early detection. When multiple inputs are involved, simple concatenation of layers might undermine few inputs' significance on the model. We employ cross-stitch mechanism which provides a way to find the optimal combination of model parameters that are used to pass the inputs to various sections of the network. Attention mechanisms have the ability of 'attending to' particular parts of the input when processing the data, allowing Cross-SEAN to be capable of representing the words which are being concentrated on, for a given tweet text.

We compare Cross-SEAN with seven state-of-the-art models for fake news detection. Experimental results show that Cross-SEAN achieves 0.95 F1 Score on CTF, outperforming seven baselines by at least 9%. We show comparative evaluation of baselines with Cross-SEAN on various features and present a thorough ablation study of Cross-SEAN to understand the importance of different features and various components of the objective function.

Our contribution III: Chrome-SEAN. For easy and real-time usage by Twitter users, we finally introduce a chrome extension, called Chrome-SEAN which uses Cross-SEAN to classify a tweet while in the tweet page. To evaluate Chrome-SEAN, we collect feedback from human subjects. We further perform online learning conditioned on the feedback and the confidence of model. The extension is deployed and configured to handle concurrent requests.

In summary, our major contributions are four-fold:

- CTF, the first labelled COVID-19 misinformation dataset.
- **Cross-SEAN**, a model to curb COVID-19 fake news on Twitter. It is one of the few semi-supervised models introduced for the task of fake news detection.
- Detailed analysis of the dataset to unfold the underlying patterns of the COVID-19 related fake tweets.
- **Chrome-SEAN**, a chrome extension to flag COVID-19 fake news on Twitter.

Reproducibility: We have made the code and the CTF dataset public at https://github.com/williamscott701/Cross-SEAN. Section 6 describes more about the settings to reproduce the results.

2. Related work

As our work revolves around fake news and semi-supervised learning, we present the related work in two parts: (i) fake news detection, and (ii) text-based semi-supervised learning. Due to the abundance of literature in both these areas, we focus our attention to those studies which we deem as pertinent to the current work.

Fake news detection: Fake news or misinformation on social media has gained a lot attention due to the exponential usage of social media. Some of early studies tried to detect fake news on the basis of linguistic features of text [15–17]. A group of recent approaches have used temporal linguistic features with recurrent neural network (RNN) [18] and modified RNN [19,20] to detect fake news. Hybrid approaches by Kwon et al. [21] combined user, linguistic, structural and temporal features for fake news classification. Lately, convolution networks have been adopted along with recurrent networks to detect fake news [22, 23]. Malhotra and Vishwakarma [24] used the graphical convolutional networks and transformer-based encodings for the task of rumour detection of tweets. They leveraged the structural and graphical properties of a tweet's propagation and tweet's text. Since satire can also lead to spread of misinformation, Rubin et al. [25] proposed a classification model using 5 features to identify satire and humour news. Another study focused on detecting fake news using n-gram analysis through the lenses of different feature extraction methods [26]. Granik and Mesyura [27] detected fake news using Naive Bayes classifier and also suggested potentials avenues to improve their model. Ozbay and Alatas [28] proposed a combination of text mining techniques and supervised artificial intelligence algorithms for the task of fake news detection. They showed that the best mean values in terms of accuracy, precision, and F-measure are obtained from the Decision Tree algorithm. Apart from textual features, visual features have also been employed for fake news detection. [29] proposed a similarity-aware fake news detection method which utilises the multi-modal data for effective fake news detection. On the similar lines, Varshney and Vishwakarma [30] developed a click-bait video detector which is another prevalent form of online false content. Despite the success of supervised models, news spreads on social media at very high speed when an event happens, only very limited labelled data is available in practice for fake news detection. Some studies such as [31,32] have been involved around weakly supervised learning for fake news detection. In similar directions, Yu et al. [33] used constrained semi-supervised learning for social media spammer detection, while Guacho et al. [34] used tensor embeddings to form a semi-supervised model for content based fake news detection. Dong et al. [35] proposed a two-path deep semi supervised learning for timely detection of fake news. They verified their system on two datasets and demonstrated effective fake news detection. Vishwakarma et al. [36] analysed the credible web sources and proposed a reality parameter for effective fake news prediction. Varshney et al. [37] developed an automated system Hoax-News Inspector for real time prediction of fake news. They used content resemblance over web search results for authenticating the credibility of news articles. Recently, Patwa et al. [38] prepared an English COVID-19 fake news dataset [39] and a Hindi hostile post dataset [40]. A few recent studies [8,41,42] have provided extensive literary surveys by investigating datasets, features and models along with potential future research prospects for fake news detection.

Semi-supervised models for text classification: Semi-supervised learning (SSL) is proved to be powerful for leveraging unlabelled data when we lack the resources to create large-scale labelled dataset. Prior research on semi-supervised learning can broadly be divided into three classes — multi-view, data augmentation and transfer learning [43]. The objective of multi-view approaches is to use multiple views of labelled as well as unlabelled data. Johnson and Zhang [44] obtained multiple views for text categorisation by learning embedding of small text regions from unlabelled data and integrating them to a supervised model. Gururangan et al. [45] and Chen et al. [46] leveraged variational autoencoders in the form of sequence-to-sequence modelling on text classification and sequential labelling. Data augmentation approaches involve augmenting either the features or labels. Nigam et al. [47] classified the text using a combination of Naive Bayes and Expectation Maximisation algorithms and demonstrated substantial performance improvements. Miyato et al. [48] utilised adversarial and virtual adversarial training to the text domain by applying perturbations to the word embeddings. Chen et al. [49] introduced MixText that combines labelled, unlabelled and augmented data for the task of text classification. They interpolated text in hidden space using Mixup [50] to create a large number of augmented training samples. Xie et al. [51] used advanced augmentation methods (RandAugment and back-translation) to effectively noise unlabelled examples. Transfer learning approaches aim to initialise task-specific model weights with the help of pre-trained weights on auxiliary tasks. Dai and Le [52] used a sequence autoencoder, which reads the input sequence into a vector and predicts the input sequence again to use unlabelled data for improving sequence learning with recurrent networks. Hussain and Cambria [53] employed a semi-supervised model based on the combined use of random projection scaling, and support vector machines to perform reasoning on a knowledge base. They showed a significant improvement in emotion recognition and polarity detection tasks over the state-of-the-art methods. Howard et al. [54] proposed the Universal Language Model Fine-tuning (ULMFiT), which has been proved as an effective transfer learning method for various NLP tasks. Both studies [52,54] showed the improvement in the performance of text classification using transfer learning.

The most of the aforementioned methods for fake news detection are tested on datasets with high volume of labelled data. Moreover, when multiple features are considered, their optimal combination is not explored. There is no published work related to COVID-19 fake news detection. We strive to address these issues by first introducing the novel CTF dataset and then leveraging the unlabelled data in order to reduce the vast dependency on the labelled data in our proposed Cross-SEAN model. We also employ cross-stitch for optimal combination of inputs into various sections of the model and show interesting analysis.

3. Dataset collection and annotation

In this section, we introduce our novel dataset, called CTF (**C**OVID-19 **T**witter **F**ake News). The formation of this dataset underwent four stages mentioned below.

Stage 1. Segregating COVID-19 related tweets: Multiple COVID-19 Twitter datasets (unlabelled) have recently been made public on Kaggle and other sources; among them, we used the datasets released by [55,56], and [57]. Alongside, there exist a few publicly available datasets containing COVID-19 related tweet IDs being released everyday in chronological order. We collected the tweet IDs from [58] and [59]. Due to the hydrating process

Table 1

Different attributes including keywords, hashtags, and sources of statements and URLs along with the respective number of tweets they are responsible for. The table compiles the numeric details of Section 3. Here, WHO: World Health Organisation, CDC: Centers for Disease Control, NIH: National Institute of Health, CPHO: Central Public Health Office, PHE: Public Health England, HHS: Human and Health Services.

Dataset collection and labelling							
Using hashtags and keywords			Using statements and tweets from organisations			Using URLs	
Major keywords	No. of tweets		Major sources	No. of tweets		Major services	No. of tweets
	Fake	Genuine		Fake	Genuine		Fake	Genuine
bioweapon	4978	0	WHO	3395	4700	Snopes	1696	1650
vaccine	3620	221	CDC	1649	2195	PolitiFact	1484	2250
trump	2874	439	NIH	2231	1705	FactCheck	1060	1500
china	2677	515	CPHO	582	470	TruthOrFiction	1042	1895
WHO	493	4018	PHE	391	425	–	–	–
at home	0	4552	HHS	405	2255	–	–	–

(which is time consuming) and the non-existence of fake tweets (as Twitter deletes them upon identification), the tweet IDs did not turn out to be very useful. However, we still considered them in our dataset to learn the language semantics explained in the subsequent sections. We also collected tweets using the Twitter API based on some predefined hashtags (e.g., 'WHO', 'covid19', 'wuhan', 'bioweapon', etc.). Since the genuineness of news correlates to the credibility of the source, we collected tweets published by the aforementioned governmental health organisations and gathered their official Twitter IDs. We extracted tweets from these accounts and considered them genuine.

Stage 2. Collecting COVID-19 supporting statements: There exist fact checking sites which analyse popular news across social media and label them as fake or genuine based on verified sources. We crawled various fact-checking sites such as *Snopes*, *PolitiFact*, *FactCheck* and *TruthOrFiction* for content related to COVID-19. We extracted URLs, the content of URLs and their corresponding labels (genuine or fake) from the fact checking websites. To support this data, more genuine URLs were extracted from the Twitter accounts of the official health bodies. To increase public awareness about any widely accepted misinformation, governmental bodies across the world have setup specific web pages,[4,5,6] which are also scraped. This stage resulted in a bulk amount of data related to the content and URLs which are known to be fake/genuine and act as the supporting statements for the next stage.

Stage 3: Filtering genuine and fake tweets: We assumed that when a fake or genuine URL is being shared, all the tweets accompanying the URL also belong to the same class as URLs are generally added in support to the text. Based on this, a total of 5.3K and 7.5K tweets were labelled as fake and genuine, respectively. Although this assumption may garner some unwanted noise since a tweet might contradict the opinion presented to the referred URL, on manual inspection we found out that this assumption surprisingly held true for most of the cases, as elaborated in the next section. In addition, all the tweets posted by governmental health organisations related to COVID-19 with specific hashtags as mentioned above, form a majority of our genuine data. This is based on the assumption that such health organisations post content which either curb fake news or are genuine in itself. We gathered 10K genuine tweets via this method. Next, we used the pre-processed tweet texts with two Transformer models, BERT [60] and RoBERTa [61], to populate the dataset further. BERT is used to generate embeddings of both tweet text and the supporting statements collected and cosine distance is computed with a high threshold of 0.9 to label the tweet into genuine or fake based on the polarity. This step

resulted in 9.7K tweets labelled as fake. For RoBERTa, we used the fine-tuned version on the Stanford Natural Language Inference (SNLI) Corpus [62]; this allowed to take in a pair of sentences and check if they are contradicting, neutral or entailing. We formed pairs of tweets and supporting statements to identify genuine or fake tweets based on contradicting and entailing results. This approach gave us an extensive set of 10.6K fake tweets.

Stage 4: Human annotation: We performed manual verification of a part of 45,261 labelled tweets (26,706 fake, 18,555 genuine) obtained from Stage 3. We employed three human annotators, who are experts in social media and have significant expertise in fact verification, to verify the labels. The annotators ended up annotating 16,000 tweets (8000 fake and 8000 genuine) with an inter-annotator's agreement of 0.82 (Krippendorf's α) with the following instructions provided:

- A tweet is considered to be 'fake' if and only if:
  - It contradicts or undermines facts from a pre-defined list. Note that a combined list was made from the aforementioned genuine sources.
  - It supports or elevates a commonly identified misinformation.
  - It is written in the form of sarcasm or humour, but promotes a misleading statement.
- Other tweets which do not satisfy any of the above, would be either unlabelled or genuine, as per the annotator's discretion.
- If the tweet text in itself does not provide enough context to annotate with confidence, the annotators could refer to the tweet and user features.

On further observation, it is found that an average of 92% labels given by the automated techniques from Stage-3 matched the labels given by the human annotators for 16,000 samples. Thus, despite using a fully-automated and fast annotation pipeline, which allowed us to have a relatively large labelled corpus, only a noise of 8% exists.

During cross-validation, we use 20% of the human-verified tweets for testing, and remaining 80% tweets along with the unverified tweets[7] constitute the training set. We maintain the same distribution of fake and genuine tweets present in the entire dataset in both the training and test sets.

4. Dataset analysis

In Table 1, we show major keywords, statements and tweets from organisations and URLs used and the number of tweets that are labelled as fake and genuine.

---

4 https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

5 https://www.cdc.gov/coronavirus/2019-ncov/index.html.

6 https://www.coronavirus.gov/.

7 It may plant some noise in the training set which a sophisticated classifier should ignore while being trained.
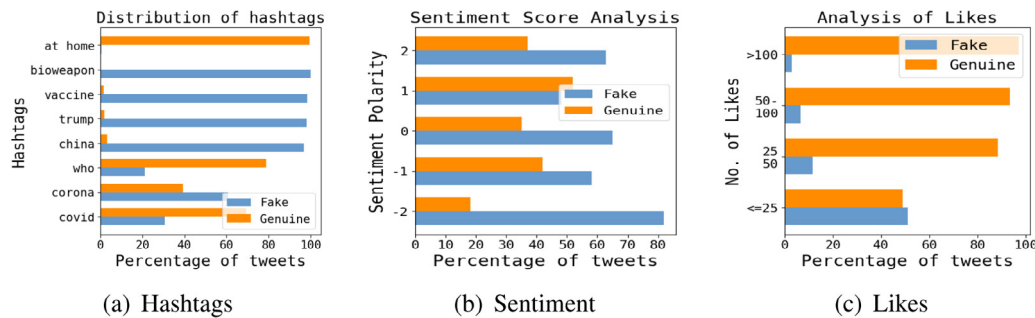
Fig. 3. (a), (b) and (c) show the distribution of hashtags, sentiment and likes across the tweets, respectively.

Presence of hashtags: Hashtags have long been an important tool on Twitter to organise, sort, follow and spread tweets. Our dataset consists of a total of 955 and 2,231 unique hashtags in genuine and fake tweets, respectively. We tabulate the distribution of hashtags for tweets in Fig. 3(a). It is evident that '#WHO' is more prominent in genuine tweets. The vast number of tweets containing '#china' and '#bioweapon' are fake tweets. Interestingly, the appearance of '#trump' hashtag in the fake tweets is much higher than the genuine tweets, pointing towards the tendency of politicisation amongst fake tweets. Even though the vaccine for COVID-19 is still under development, the recurrent use of '#vaccine' in fake tweets may suggest the tendency of spreading rumours with false remedies. The dominance of hashtags such as '#togetherathome', '#stayhome' and '#socialdistancing' in the genuine tweets suggests that they might have been used to spread positive social messages.

Presence of URLs: To account for prevalence of misinformation, we analyse the URLs present in our entire dataset. A total of 14,830 genuine and 8,761 fake tweets contain at least one URL, thus averaging to 0.87 and 0.35 URLs per genuine and fake tweet, respectively. The contrast between the numbers may suggest that in general, genuine tweets have a higher tendency of supporting the claims.

Presence of multimedia: Twitter supports three types of media formats in a tweet– photo (P), video (V) and GIF (G). However, it supports only one type of media in a particular tweet with a limit of four photos and only one video/GIF. In our dataset, fake tweets contain a total of 2,491 media files (2036P, 381 V, 74G) across 2,344 tweets, with an average of 0.0988 per tweet, while genuine tweets contain 1,473 media (1129P, 339 V, 5G) with an average of 0.0834.

Sentiment of tweets: To obtain overall sense of public opinion related to COVID-19, we analyse sentiment of the tweets [63] using the texblob[8] tool. Fig. 3(b) shows that in the highly negative (-2) and neutral (0) sentiment zones, fake news are grouped more than the genuine news. The average sentiment polarity for fake tweets is 0.05 compared to 0.096 in genuine tweets, on a scale of -2 to 2, as shown in Fig. 3(b). [Similar results were also obtained from latest state-of-the-art polarity classification methods for long reviews and short tweets [64–66]].

Likes and retweets: The existing propagation based approaches [27,67] showed the significance of likes and retweets for fake news detection. The average number of likes per genuine tweet is found to be 142.65, which is significantly higher than that (4.25) of fake tweet. The tweet-wise data of likes is summarised in Fig. 3(c). The large number tweets of popular public health organisation explains the higher average likes per genuine tweet. About 64% of fake tweets in our dataset are retweets of some

other tweet, 8% of the fake retweets are quoted with the comments, and 35% of genuine tweets are retweets with 8% of them being retweets with comment.

Visual representations: We show t-SNE visual representations of labelled and unlabelled tweets on tweet text, tweet features and user features in Figs. 3–5. Fig. 3 shows tweet text representations on labelled and unlabelled data. Sentence BERT is used to convert the tweet text to vector form. While the overlap of genuine and fake tweets can be observed from Fig. 4(a), the polarisation of topics can be observed from the unlabelled data from Fig. 5(a). Certain user features and tweet features are identified and are mentioned in Section 5.1; these are in turn used for the visualisations on labelled and unlabelled data in Figs. 4 and 5 respectively. The polarisation in Fig. 5 supports the same in Fig. 5(a). The labelled representation shows high non-linear overlap and indicates the complexity of the classification task.

5. Cross-SEAN: Our proposed method

In this section, we describe Cross-SEAN[9] for fake news detection. We explain individual components of the model, followed by the training strategy. Fig. 6 shows the architecture of Cross-SEAN and Fig. 7 shows the cross-stitch unit.

5.1. Explicit tweet and user features and external knowledge

Monti et al. [68] showed that content, social context or propagation in isolation is insufficient for neural models to detect fake news. Hence, we employ additional features related to both the users and tweets along with the content of the tweets. For the tweet features (TFs), we consider the attributes available in the tweet object and some handcrafted features from the tweet, amounting to a total of 10 features – *number of hashtags, number of favourites, number of retweets, retweet status, number of URLs present, average domain score of the URL(s), number of user mentions, media count in the tweet, sentiment of the tweet text, counts of various part-of-speech tags* and *counts of various linguistic sub-entities*. Polarisation of users on similar beliefs is widely observed on Twitter [69]. To capture this, we extract 8 features for each corresponding user (UFs) – *verified status, follower count, favourites count, number of tweets, recent tweets per week, length of description, presence of URLs* and *average duration between successive tweets*.

These features can provide additional information of the user characteristics and their activities. These not only help the model identify bots and malicious fake accounts, but also help recognise a pattern amongst users who post false and unverified information.

On visualising the tweet and user features on labelled and unlabelled data in Fig. 3, we observe the formation of clusters

---

[9] **Cross**-Stitch based **S**emi-Supervised **E**nd-to-End **A**ttention **N**eural Network.

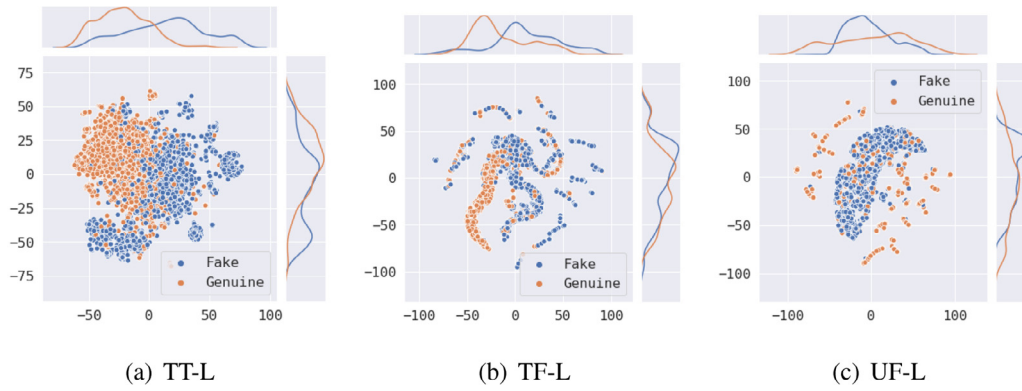(a) TT-L   (b) TF-L   (c) UF-L

Fig. 4. (a), (b) and (c) show the t-SNE visual representations of tweet text, tweet features and user features of the labelled data, respectively. Here, TF → Tweet Features, UF → User Features and L → Labelled Data.

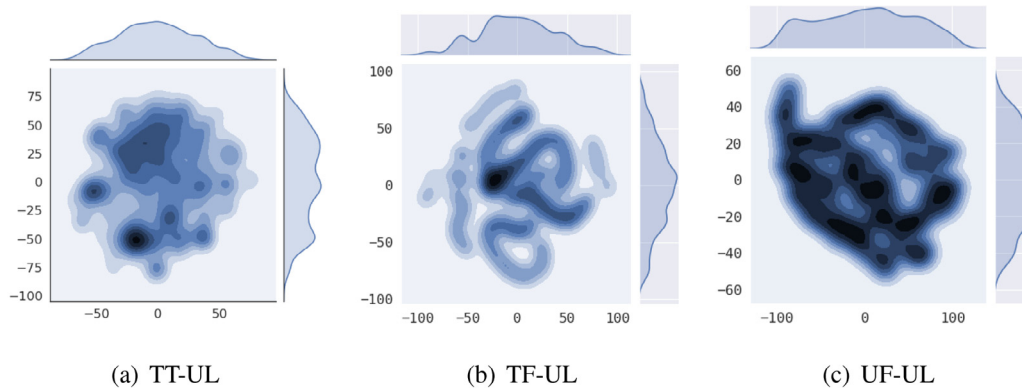

(a) TT-UL   (b) TF-UL   (c) UF-UL

Fig. 5. (a), (b) and (c) show the t-SNE visual representations of tweet text, tweet features and user features of the unlabelled data, respectively. Here, TF → Tweet Features, UF → User Features and UL → Unlabelled Data.

of similar tweets, indicating the polarity of the tweets. From Fig. 4(c), we also observe that users posting fake tweets tend to form a cluster, and users posting genuine tweets are scrambled across the whole feature space. Few features of user posting genuine tweets are highly similar to the features of users posting fake tweets, thus overlapping with the fake tweet cluster. From Fig. 5, the unsupervised user features show the dense polarity across the whole latent space, while the tweet features are wide spread, showing the diverse set of attributes in our unlabelled data. These features are further used in the classification.

Feature based neural models learn a generalised function from a limited manifold of the training data, and thus have a tendency to perform poorly when the topics are variant. To overcome this in Cross-SEAN, we use external knowledge, for the content relating to tweet text, as an input to the model. We use classical text processing techniques to find a shortened contextual form of the tweet text and use it as a query to retrieve the top Google Search results, sorted in accordance to relevance [24,36,37]. From each web-article returned from the search, a particular number of text sentences are retrieved which are the closest to the original tweet text, as measured using cosine similarity of the BERT Sentence Embeddings [60] of the two. This is done until k (=10, by default) sentences are retrieved for the tweet.

In addition to this, we make use of the large amount of unlabelled data (21.85M) available in CTF–

- We use one-half of the unlabelled data to fine-tune word embeddings to encode the tweet text. We expect this to help the model learn the linguistic, syntactic and contextual composition of not only general Twitter Data but also the domain data, i.e., the COVID-19 pandemic in case of CTF.

- We leverage the other half of the unlabelled data for unsupervised training using an additional adversarial loss. Experimental results presented in Section 6.2 show that doing this reduces stochasticity and makes the model more robust with the nature of adversarial training.

We elaborate on various components of the model architecture and the training intricacies in the following sections.

5.2. Model architecture

Our entire training data is composed of labelled and unlabelled samples, denoted by X_L and X_U respectively. X_L consists of a total of n_L data points: $(x_L^1, y_L^1), (x_L^2, y_L^2), \cdots, (x_L^{n_L}, y_L^{n_L})$, where x_L^i is the ith tweet and y_L^i is its label. X_U consists of a total of n_U unlabelled data points: $x_U^1, x_U^2, \ldots, x_U^{n_U}$. In both the cases, each input sample, x_K^i (for $K \in (L, U)$) comprises four input sub-sets − tweet text (x_{TT}^i), external knowledge text (x_{EK}^i), tweet features (x_{TF}^i) and user features (x_{UF}^i).

In each pass through our model, these four inputs are encoded separately as described below.

Encoding textual data: The tweet text of sequence length N is represented as a one-hot vector of vocabulary size V. A word embedding layer $E \in R^{V \times D}$ transforms the one-hot vector into a dense tensor $e \in R^{N \times D}$ consisting of $(e^1, e^2, \ldots, e^N)$. These token vectors are further encoded using a Bidirectional LSTM, the forward and backward layers of which process the N vectors in opposite directions.

The forward LSTM emits a hidden state h_{ft} at each time-step, which is concatenated with the corresponding hidden state h_{bt} of
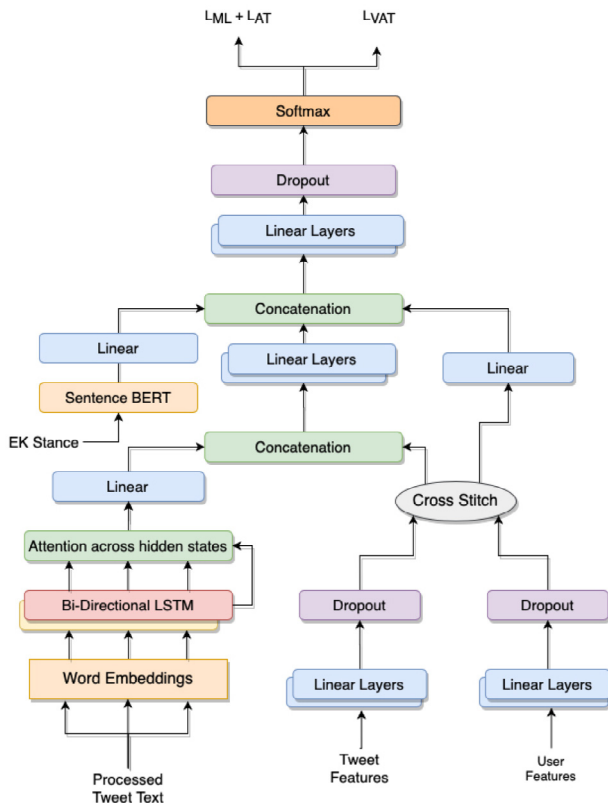
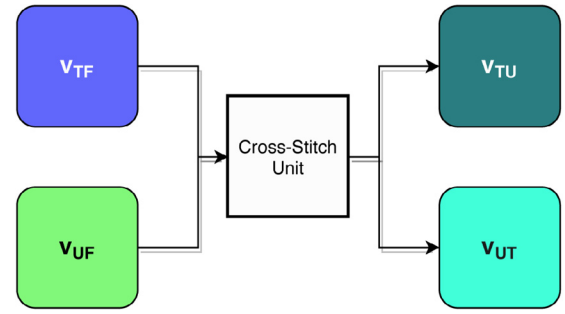**Fig. 6.** A schematic diagram of Cross-SEAN.

Fig. 7. Working of a cross-stitch unit. Here, the notation is as defined in Eq. (5). Note that the weights of the linear layers in the Cross-stitch unit are initialised with a unit matrix.

the backward LSTM to produce a vector $h_t \in R^{(2 \times H)}$,

$$h_t = h_{ft} \oplus h_{bt}, \forall t \in [1, N] \tag{1}$$

where H is the hidden size of each LSTM layer.

At each layer, a final state output $f_k \in R^H$ is also obtained ($\forall$ $k \in (f, b)$).

At this stage, a net hidden vector h containing N hidden vectors from the two LSTM layers is combined with the final state vector f using attention across the hidden states, given as:

$$v = \sum_{j=1}^{N} \alpha_{ij} h_j; \alpha_{ij} = Softmax(h_i \bullet f_j) \tag{2}$$

where,

$$f = f_f \oplus f_b, f \in R^{2 \times H} \tag{3}$$

$$h = h_1 \cdots \oplus \cdots h_N, h \in R^{N,(2 \times H)} \tag{4}$$

We refer vector v obtained after attention across the hidden states as v_{TT}, representing the encoded feature of the tweet text.

In addition to this, we use Sentence BERT [60] to find contextual embedding e_{EK} of the external knowledge corresponding to each input batch. We do this considering the vast difference between our tweet text input and the external knowledge text. The e_{EK} vector is then passed through a linear layer to obtain an encoded representation v_{EK} of the external knowledge.

Encoding tweet and user features: As shown in Fig. 3, we follow a highly concurrent yet distinct mechanism to encode both tweet and user features. Firstly, $x_{TF} \in R^{K_t}$ and $x_{UF} \in R^{K_u}$ are passed through separate linear layers which interpolate them to higher dimensional dense feature vectors $v_{TF} \in R^{K_T}$ and $v_{UF} \in R^{K_U}$, respectively. As both x_{TF} and x_{UF} are handcrafted, we employ cross-stitch units, which not only allow the model to learn the

best combination of inputs from both the features and share across multiple layers in the network, but also introduce a common gradient flow path through the non-linear transformation. The transformation produced by cross-stitch is as follows:

$$v'_j = \alpha_{ij} \bullet v_j + \beta_i, \forall i, j \in (1, K_T + K_U) \tag{5}$$

where α_{ij} and β_i denote the weights of the fully connected layer performing the cross-stitch operations.

The two outputs of the cross-stitch are denoted by v_{TU} and v_{UT}, respectively.[10] Note that the shape of the two vectors remains unchanged after this transformation.

Connected components in Network: We concatenate v_{TT} and v_{TU}, which are the transformed feature vectors of the tweet text x_{TT} and tweet features x_{TF}, respectively. This produces $v_T = v_{TT} \oplus v_{TU}$, a concatenated representation of all textual features. This is done considering the inherent similarity between the tweet text and the tweet features over user features. We then perform affine transformations of the three vectors, v_T, v_{EK} and v_{UT}, through separate feed-forward linear layers and concatenate to obtain the final decoded vector v, effectively containing transformed feature representations from all the inputs. The vector v is then down-scaled using a fully-connected network, regularised using dropout before finally obtaining the probability distribution across the two classes.

$$p(y|x; \theta) = Softmax(v') = Softmax(||v'_T||v'_{UT}||') \tag{6}$$

where, v' represents the transformed vector after it passes through the respective feed forward sub-network, and θ represents the model parameters at the current time (from now on, we refer to this as $f(x)$).

5.3. Training strategies

For training our model, we use a mixed objective function, which is a weighted sum of both supervised and unsupervised losses:

$$L_{mix} = \lambda_{ML} L_{ML} + \lambda_{AT} L_{AT} + \lambda_{VAT} L_{VAT} \tag{7}$$

The losses are as follows: (i) L_{ML} represents maximum likelihood loss and minimises the loss between the predicted and true labels. (ii) Additionally, we use the Adversarial Training Loss L_{AT}, which introduces a regularisation with model training by adding a denoising objective [48]. The goal through this training is to make the model robust to adversarial perturbations in the input. We find this specially useful for fake news detection as it allows the model to attend to a wide spectrum of tweets with minor

---

[10] The first letter in the subscript of v denotes feature vector assuming that it contains most information from the same vector.

Table 2
Features used by the competing models and performance comparison on CTF (TT: Tweet Text, TF: Tweet Features, UF: User Features, UL: Unlabelled Data).

Model	Features used by the model				Performance			
	TT	TF	UF	UL	Accuracy	Precision	Recall	F1 Score
MTL	✓	✓			0.79	0.77	0.82	0.79
1HAN	✓				0.89	0.60	0.87	0.71
16HLT-HAN	✓				0.87	0.68	0.86	0.76
3HAN	✓			✓	0.89	0.77	0.82	0.80
CSI	✓		✓		0.87	0.80	0.91	0.85
dEFEND	✓		✓		0.89	0.83	0.89	0.86
MixText	✓			✓	0.87	0.83	0.84	0.84
Cross-SEAN	✓	✓	✓	✓	**0.954**	**0.946**	**0.961**	**0.953**

variations to improve the generality. An adversarial signal r_{adv}, defined in terms of the L_2 norm on the gradient g_L, with current model parameters is used to perturb the word embedding inputs e of x_{TT}, $e\star = e + r_{adv}$, even when this perturbation depends upon the gradient computed over the output w.r.t all the labelled inputs x_L. The L_{AT} objective function in Eq. (10) is given as a modification of L_{ML} (Eq. (8)). (iii) It can be observed that the above two objectives require us to know the true label of the data input, thus pertaining to the labelled data only. Here, to expand the concept of adversarial training to unlabelled data, we make use of virtual adversarial training loss L_{VAT}, which too is aimed to add robustness against adversarial inputs. Just as in Eq. (10), we apply the perturbation on the word embedding e, except r_{adv} is now defined as in Eq. (13). δ represents a small random perturbation vector [48], using a 2nd-order Taylor series expansion followed by the power iteration method. The VAT loss is then defined as in Eq. (14). We denote $f(x) = h(E(x))$, where $E(x) \in R^{N \times D}$ is the word embedding vector.

$$L_{ML} = \frac{-1}{n_L} \sum_{i=1}^{n_L} y_i \log(f(x_i)) + (1 - y_i) \log(1 - f(x_i)) \qquad (8)$$

$$r_{adv} = -\epsilon_L / \|g_L\|_2; \; g_L = -\nabla_{x_L} \log(f(x_L)) \qquad (9)$$

$$L_{AT} = \frac{-1}{n_L} \sum_{i=1}^{n_L} P + Q \qquad (10)$$

where,

$$P = y_i \log(h(E(x_i)) + r_{adv}) \qquad (11)$$

$$Q = (1 - y_i) \log(1 - f(h(E(x_i) + r_{adv}))) \qquad (12)$$

$$r_{v-adv} = \epsilon g / \|g\|_2; \; g = -\nabla_x KL[f(x) \parallel h(E(x) + \delta)] \qquad (13)$$

$$L_{VAT} = \frac{1}{n_L + n_U} \sum_{i=1}^{n_L + n_U} KL[f(x) \parallel h(E(x) + r_{v-adv})] \qquad (14)$$

6. Experimental setup and results

All our experiments were performed on a single 16 GB Nvidia Tesla V-100 GPU. Our base model is a single layer Bi-LSTM with a maximum sequence length of 128 and a hidden dimension of 512. We performed experiments with a wide range of embedding sizes ranging from 128 to 768 and found the best results with 300 dimensions. We initially fine-tuned the word embeddings on $\sim 10M$ unlabelled tweet texts before using them for training. We used the *Adam* optimiser for all our experiments with a learning rate of 0.001, $\beta_1 = 0.90$, $\beta_2 = 0.98$ and a decay factor of 0.5. We used dropout with p_{drop} of 0.3 in all our feed-forward networks, where the number of layers exceeds 2. Early stopping with a patience of 20 was also used along with gradient clipping with a maximum L_2 norm of 1. We kept λ_{ML}, λ_{AT} and λ_{VAT} as 1.

Table 3
Results of Cross-SEAN with different variations of the mixed objective function.

Objective function			Result	
ML	AT	VAT	Accuracy	F1 Score
✓			0.910	0.907
✓	✓		0.936	0.930
	✓	✓	0.854	0.860
✓		✓	0.936	0.930
✓	✓	✓	**0.954**	**0.953**

6.1. Comparative evaluation

We compare Cross-SEAN with seven state-of-the-art methods described as follows. **MTL** [70] uses a multitask learning framework by leveraging soft parameter sharing on classification (primary) and regression (secondary) tasks based on tweet text and tweet features. 1HAN and 3HAN [71] use hierarchical attention based GRU networks. **1HAN** is the base version of 3HAN, where **3HAN** uses 3-level hierarchical attention for words, sentences and headlines learning in a bottom up manner. **16HLT-HAN** [72] uses hierarchical structure by applying attention mechanism at both word and sentence levels. **CSI** [73] uses a three module approach that consists of *Capture*, *Score* and *Integrate*, combining what they define as the three common characteristics among fake news, i.e., *text*, *response* and *source* to identify misinformation. Furthermore, we also use **dEFEND** [74] as a baseline, which uses a GRU-based word-level and sentence-level encoding along with a module for sentence-comment co-attention. **MixText** [49] is a semi-supervised approach that produces results by leveraging large amount of training samples and interpolating text in hidden space.

Table 2 shows that Cross-SEAN outperforms all the baselines by a margin of at least more than 6% accuracy and 9% F1 Score, with dEFEND being the best baseline.

6.2. Ablation study

(a) Objective functions: In Table 3, we test the performance of Cross-SEAN on different combinations of the mixed objective function. We vary the values of λ_{ML}, λ_{AT}, λ_{VAT} between 0-1. A steady increase in the performance can be seen as we move from a vanilla supervised training objective (only maximum likelihood loss) to an additional semi-supervised mixed objective function.

Fig. 8 shows the variation of different objectives functions – ML, AT and VAT, individually, when trained with different combinations of the mixed objective function. For instance, Fig. 8(a) shows the variation of the individual ML Loss when different combinations of the net objective function is used.

From Fig. 8(a), the regularisation effect of the two adversarial losses, AT and VAT, is apparent as it can be observed that their introduction considerably effects the individual ML loss, making

Table 4

Results with various fully-connected network combinations. Here, TF*i* and UF*i* represent the *i*th layer transposing a feature vector of tweet features and user features respectively. Two joined cells represent a concatenated form of the respective vectors feeding as inputs to the corresponding layer.

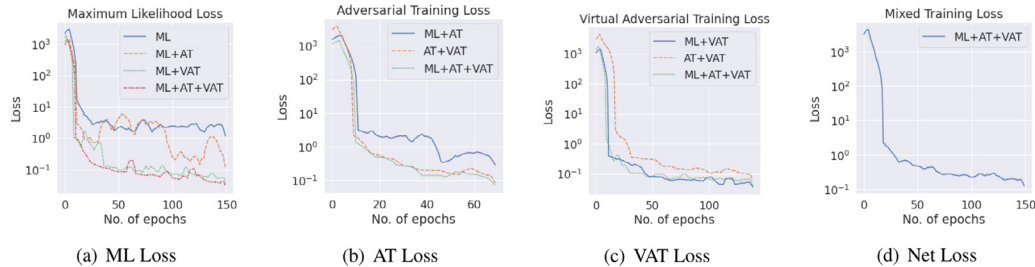| Linear layers								Attention	Cross-stitch	Performance	
TF1	UF1	TF2	UF2	TF3	UF3	TF4	UF4			Accuracy	F1 score
	128									0.910	0.884
	64		256					✓		0.932	0.935
64	64	256	256					✓	✓	0.931	0.934
	64		256		256		512			0.939	0.942
	64		256		256		512	✓		0.927	0.944
64	64	256	256		256			✓	✓	**0.954**	**0.953**

Fig. 8. Variation of individual loss functions of Cross-SEAN with different combinations of the mixed objective function.

it drop to a larger extent, in fewer iterations. Even though the introduction of AT alone seems to make the loss curve more stochastic, the net loss is considerably lower. This can be seen in addition to the surprising *smoothing* effect which is observed wherever the VAT loss is considered, including Figs. 8(b) and 8(c). These two properties of AT and VAT losses respectively, motivate their usage together, thus resulting into an efficient and smooth decrease of loss and strengthening our hypothesis of leveraging unlabelled data. This is further ensured by another interesting observation by using only AT and VAT losses for the training – although as expected, we achieve a deteriorated accuracy as shown in Table 3, the corresponding losses in Figs. 8(b)–8(c) show high consistency and smoothness. Fig. 8(d) shows the final loss curve when all the 3 losses are used, i.e., when $\lambda_{ML} = \lambda_{AT} = \lambda_{VAT} = 1$.

(b) Model Components: Table 4 shows the importance of different components used in Cross-SEAN such as cross-stitch, attention and feed-forward layers for tweet and user features. We experiment across several combinations of tweet features and user features with concatenation and usage of cross passing through various layers as shown in Table 4. We find that the best architecture is with the cross-stitch on tweet and user features when one output of the cross-stitch is combined in the early stages of the network and the other output is fused in the later stage. Also the use of attention shows performance improvement of the final model.

In our initial set of experiments, the cross-stitch was introduced between the encoded representation of the tweet text, obtained after passing it through Bi-LSTM, and a concatenated form of tweet and user features. A considerable difference in the performance is observed between the two, the former being the superior one. We relate this to the fact that the encoded representation of the tweet text is considerably different from the additional features, while they in themselves are very similar. Further, since the tweet features are inherently more similar to the tweet text, the cross-stitch output corresponding to the tweet features is first concatenated to the encoded tweet text and lastly with the user features. This is also shown in Table 4, where the architecture used in the last row evidently outperforms the one in the 3rd row, which represents concatenation of the three outputs on the same level.

7. Chrome-SEAN: A chrome extension

Cross-SEAN is an end-to-end model which enables for identification of fake tweets in real time. Keeping the users warned is a very important step and would help with an easy access through the browser. In order to help users detect misinformation on Twitter in real time, we deploy Cross-SEAN as a Chrome browser extension, called Chrome-SEAN that replicates the performance of the model while performing a lot of other features as well.

Chrome-SEAN is built as Chrome extension, which uses jQuery[11] to send and receive requests from POST API method. We deployed the Cross-SEAN model using Flask[12] in our local servers which can receive the POST API requests concurrently. To handle the load balancing over multiple concurrent requests, we use Redis.[13] The server is not burdened with resource intensive requests, and the combination of Flask and Redis performs efficient communication through APIs.

Chrome-SEAN first identifies the tweet ID through the URL while scanning Twitter, and sends it to the server using an API. Chrome-SEAN also provides the option to enter the tweet ID manually. Upon requesting to Cross-SEAN, the raw data is first transformed to the necessary format and then passed through the model. The detected class along with its confidence from the softmax layer is returned back to the extension and displayed. Fig. 9 shows the working of Chrome-SEAN in two stages. In the former stage, the extraction of the tweet is performed in the browser side and is instant, whereas in the latter stage, verification of the tweet takes on an average of 1.2 s per tweet (single API request).

As shown in Fig. 9, we take users' feedback on our final classification output and consider it as a true label in the extended online dataset. Additionally, we employ an online training mechanism on the basis of users' feedback if it differs from the class identified and check the confidence of the model; the model is trained only if the confidence is lower than 0.6. We take special care before online training to make the model robust to attackers

---

[11] https://jquery.com.

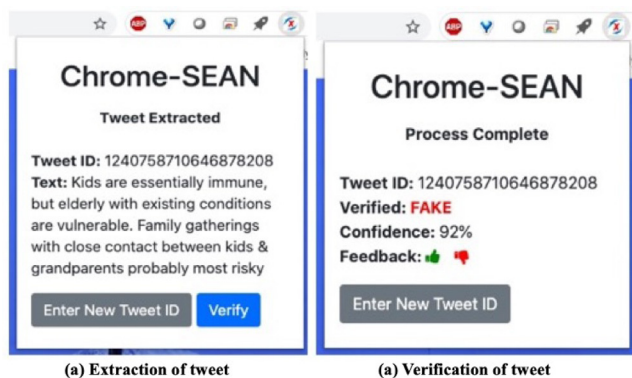[12] https://flask.palletsprojects.com/en/1.1.x/.

[13] https://redis.io/.

Fig. 9. The working of Chrome-SEAN, a Chrome extension of Cross-SEAN.

attempting to pollute the results. To handle load balancing on the server, we make use of Redis.

User Study: Chrome-SEAN was tested by 35 users until now. We first randomly sampled tweets from the human-annotated set of tweets which were not a part of the training set, assigned them to users and asked them to test on similar tweets, totalling 215 tweet inputs, ranging from a wide variety of sub-topics, users and timelines. It was observed that 67% of these input tweets were made within the last 7 days, 53% were from new users with less than 5 tweets, and 85% had a retweets count of less than 10.

We asked users to provide feedback on each tweet they tested with Chrome-SEAN, in accordance with the true label. We found that 203 out of 215 ratings were positive, i.e., deeming the prediction by Chrome-SEAN correct, resulting into an accuracy of 94% and F1 Score of 94.3%. Such high level of accuracy on such a diverse set of inputs depicts Cross-SEAN's ability to pick the appropriate input features when making a prediction.

8. Discussion and conclusion

This work introduced the task of COVID-19 fake news detection on Twitter. We collected related tweets from diverse sources. Post human annotations, we proposed CTF, the first labelled Twitter dataset, consisting of COVID-19 related labelled genuine and fake tweets along with a huge set of unlabelled data. We also presented a thorough analysis to understand surface-level linguistic features.

As the amount of labelled data is limited, we made use of the vast unlabelled data to train the neural attention model in a semi-supervised fashion as learning the semantic structures of language around COVID-19 helps the model learn better. We collected external knowledge for all the tweets by taking the most relevant stance from credible sources on the web. As fake news around COVID-19 are emerging, even if the model is not trained on a certain fake news topics, we assume that external knowledge from a trusted source could help aid the classification. We built a neural attention model which takes various inputs such as tweet text, tweet features, user features and external knowledge for each tweet. We employed cross-stitch units for optimal sharing of parameters among tweet features and user features. As tweet text and tweet features are closely related, we performed optimal sharing of information by concatenating one output of cross-stitch early in the network and the other latter. Maximum likelihood and adversarial training are used for supervised loss, while virtual adversarial training for unsupervised loss. Usage of adversarial losses further adds regularisation and robustness to the model. We then incorporated this model into Cross-SEAN, a novel cross-stitch model which performs under a semi-supervised setting by leveraging both unlabelled and labelled data with optimal data sharing across various tweet information.

Cross-SEAN is highly effective, outperforming seven state-of-the-art models significantly. We contrasted features of baseline models with Cross-SEAN and showed various metrics. We showed a thorough ablation study with various fully-connected network combinations of the model and the respective accuracy contrasting the importance of individual components of the model. We also showed variation of individual loss functions with the different configurations of the mixed objective function.

To make use of Cross-SEAN in real time by general users, we developed Chrome-SEAN, a chrome extension based on Cross-SEAN to flag fake tweets, which showed reasonable performance in a small-scale user study. Chrome-SEAN is built to be robust to handle vast amount of concurrent requests. We introduced several features to Chrome-SEAN which can further help collect labelled data using user feedback. Cross-SEAN further trains in an online fashion, for a given feedback if the confidence of the model is low. Chrome-SEAN is further tested by human subjects.

Shortcomings of Cross-SEAN: We observe following shortcomings of Cross-SEAN:

- The nature of language used in micro-blogging sites such as Twitter, in certain times makes the external knowledge noisy. Often times, a few trusted news sources on the Internet are biased on political topics which in turn create bias in the external knowledge.
- Although external knowledge adds additional information relative to the test time helping emerging fake news, it may not promise complete robustness and early detection.
- Although the tweet features, user features and external knowledge can attribute to general fake news, Cross-SEAN is a model specifically tuned for COVID-19 fake news, and is not tested on general fake news on Twitter.

Future work: We plan improve on the following points:

- We intend to study the dynamic graph structure of the follower–followee and tweet–retweet network, and extract representations from tweet and user nodes to help early detection of COVID-19 fake news.
- We will add additional improved filters to the process of extracting external knowledge to remove possible bias and noise.
- We will work towards explainability of Cross-SEAN using the current structures of attention mechanism.
- We plan to incorporate semantic information from other forms of media such as images, GIFs or videos which are readily available with the tweets. Even the textual information present in such media will be extracted and used for detection.

CRediT authorship contribution statement

William Scott Paka: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing. **Rachit Bansal:** Conception and design of study, Acquisition of data, Writing - review & editing. **Abhay Kaushik:** Conception and design of study. **Shubhashis Sengupta:** Writing - original draft, Writing - review & editing. **Tanmoy Chakraborty:** Conception and design of study, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

[1] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, J. Econ. Perspect. 31 (2) (2017) 211–236.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, 2017, arXiv:1708.01967.

[3] N. Kshetri, J. Voas, The economics of "fake news". IT prof., 19, (6) 2017, pp. 8–12.

[4] A. Kucharski, Study epidemiology of fake news, Nature 540 (7634) (2016) 525.

[5] J. Golbeck, M. Mauriello, B. Auxier, K.H. Bhanushali, C. Bonk, M.A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J.B. Everett, et al., Fake news vs satire: A dataset and analysis, in: Proceedings of the 10th ACM Conference on Web Science, 2018, pp. 17–21.

[6] D. Varshney, D.K. Vishwakarma, A review on rumour prediction and veracity assessment in online social network, Expert Syst. Appl. (2020) 114208.

[7] D.K. Vishwakarma, C. Jain, Recent state-of-the-art of fake news detection: A review, in: 2020 International Conference for Emerging Technology (INCET), IEEE, 2020, pp. 1–6.

[8] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (5) (2020) 1–40.

[9] K.-C. Yang, C. Torres-Lugo, F. Menczer, Prevalence of low-credibility information on twitter during the covid-19 outbreak, 2020, arXiv preprint arXiv:2004.14484.

[10] P.N. Howard, G. Bolsover, B. Kollanyi, S. Bradshaw, L.-M. Neudert, Junk news and bots during the US election: What were michigan voters sharing over Twitter, 2017, CompProp, OII, Data Memo.

[11] K.H. Jamieson, J.N. Cappella, Echo Chamber: Rush Limbaugh and the Conservative Media Establishment, Oxford University Press, 2008.

[12] J.H. Harvey, W.J. Ickes, R.F. Kidd, New Directions in Attribution Research: Vol. 1, Psychology Press, 2018.

[13] M. Fisher, Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism, Wash. Post 23 (2013).

[14] S. Masud, S. Dutta, S. Makkar, C. Jain, V. Goyal, A. Das, T. Chakraborty, Hate is the new infodemic: A topic-aware modeling of hate speech diffusion on Twitter, 2020, http://arxiv.org/abs/2010.04377.

[15] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 675–684.

[16] V. Qazvinian, E. Rosengren, D.R. Radev, Q. Mei, Rumor has it: Identifying misinformation in microblogs, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1589–1599.

[17] A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, Tweetcred: Real-time credibility assessment of content on twitter, in: International Conference on Social Informatics, Springer, 2014, pp. 228–243.

[18] J. Ma, W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: IJCAI, AAAI Press, 2016, pp. 3818–3824.

[19] T. Chen, X. Li, H. Yin, J. Zhang, Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 40–52.

[20] S.R. Sahoo, B. Gupta, Multiple features based approach for automatic fake news detection on social networks using deep learning, Appl. Soft Comput. 100 (2021) 106983, http://dx.doi.org/10.1016/j.asoc.2020.106983, http://www.sciencedirect.com/science/article/pii/S1568494620309224.

[21] S. Kwon, M. Cha, K. Jung, Rumor detection over varying time windows, PLoS One 12 (1) (2017).

[22] Y. Liu, Y.-F. Wu, Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[23] R.K. Kaliyar, A. Goswami, P. Narang, S. Sinha, Fndnet – a deep convolutional neural network for fake news detection, Cogn. Syst. Res. 61 (2020) 32–44, http://dx.doi.org/10.1016/j.cogsys.2019.12.005, http://www.sciencedirect.com/science/article/pii/S1389041720300085.

[24] B. Malhotra, D.K. Vishwakarma, Classification of propagation path and tweets for rumor detection using graphical convolutional networks and transformer based encodings, in: 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), IEEE, 2020, pp. 183–190.

[25] V.L. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the Second Workshop on Computational Approaches to Deception Detection, 2016, pp. 7–17.

[26] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Secur. Priv. 1 (1) (2018) e9.

[27] M. Granik, V. Mesyura, Fake news detection using naive Bayes classifier, in: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), IEEE, 2017, pp. 900–903.

[28] F.A. Ozbay, B. Alatas, Fake news detection within online social media using supervised artificial intelligence algorithms, Physica A 540 (2020) 123174.

[29] X. Zhou, J. Wu, R. Zafarani, SAFE: Similarity-aware multi-modal fake news detection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2020, pp. 354–367.

[30] D. Varshney, D.K. Vishwakarma, A unified approach for detection of clickbait videos on youtube using cognitive evidences, Appl. Intell. 1–22.

[31] S. Helmstetter, H. Paulheim, Weakly supervised learning for fake news detection on Twitter, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 274–277.

[32] G. Gravanis, A. Vakali, K. Diamantaras, P. Karadais, Behind the cues: A benchmarking study for fake news detection, Expert Syst. Appl. 128 (2019) 201–213.

[33] D. Yu, N. Chen, F. Jiang, B. Fu, A. Qin, Constrained NMF-based semi-supervised learning for social media spammer detection, Knowl.-Based Syst. 125 (2017) 64–73.

[34] G.B. Guacho, S. Abdali, N. Shah, E.E. Papalexakis, Semi-supervised content-based detection of misinformation via tensor embeddings, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2018, pp. 322–325.

[35] X. Dong, U. Victor, L. Qian, Two-path deep semisupervised learning for timely fake news detection, IEEE Trans. Comput. Soc. Syst. (2020).

[36] D.K. Vishwakarma, D. Varshney, A. Yadav, Detection and veracity analysis of fake news via scrapping and authenticating the web search, Cogn. Syst. Res. 58 (2019) 217–229.

[37] D. Varshney, D.K. Vishwakarma, Hoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles, J. Ambient Intell. Humaniz. Comput. (2020) 1–14.

[38] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. PYKL, A. Das, A. Ekbal, M.S. Akhtar, T. Chakraborty, Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts, in: Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages During Emergency Situation (CONSTRAINT), Springer, 2021.

[39] P. Patwa, S. Sharma, S. PYKL, V. Guptha, G. Kumari, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, 2020, arXiv preprint arXiv:2011.03327.

[40] M. Bhardwaj, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Hostility detection dataset in hindi, 2020, arXiv preprint arXiv:2011.03588.

[41] P. Meel, D.K. Vishwakarma, Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities, Expert Syst. Appl. (2019) 112986.

[42] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Inform. Sci. 497 (2019) 38–55.

[43] D.S. Sachan, M. Zaheer, R. Salakhutdinov, Revisiting lstm networks for semi-supervised text classification via mixed objective function, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6940–6948.

[44] R. Johnson, T. Zhang, Semi-supervised convolutional neural networks for text categorization via region embedding, in: Advances in Neural Information Processing Systems, 2015, pp. 919–927.

[45] S. Gururangan, T. Dang, D. Card, N.A. Smith, Variational pretraining for semi-supervised text classification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5880–5894, http://dx.doi.org/10.18653/v1/P19-1590, https://www.aclweb.org/anthology/P19-1590.

[46] M. Chen, Q. Tang, K. Livescu, K. Gimpel, Variational sequential labelers for semi-supervised learning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 215–226, http://dx.doi.org/10.18653/v1/D18-1020, https://www.aclweb.org/anthology/D18-1020.

[47] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using EM, Mach. Learn. 39 (2–3) (2000) 103–134.

[48] T. Miyato, A.M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, 2016, arXiv preprint arXiv:1605.07725.

[49] J. Chen, Z. Yang, D. Yang, Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2147–2157, http://dx.doi.org/10.18653/v1/2020.acl-main.194, https://www.aclweb.org/anthology/2020.acl-main.194.

[50] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, 2017, arXiv preprint arXiv:1710.09412.

[51] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, Q.V. Le, Unsupervised data augmentation for consistency training, 2019, arXiv preprint arXiv:1904.12848.

[52] A.M. Dai, Q.V. Le, Semi-supervised sequence learning, in: Advances in Neural Information Processing Systems, 2015, pp. 3079–3087.

[53] A. Hussain, E. Cambria, Semi-supervised learning for big social data analysis, Neurocomputing 275 (2018) 1662–1673.

[54] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, 2018, arXiv preprint arXiv:1801.06146.

[55] Carlson, Coronavirus tweets, tweets (json) for coronavirus on kaggle, 2020, https://www.kaggle.com/carlsonhoo/coronavirus-tweets online; accessed 2020.

[56] S. Smith, Coronavirus (covid19) tweets - early april, 2020, https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april Online; accessed 2020.

[57] S. Celin, COVID-19 tweets afternoon 31.03.2020., 2020, https://www.kaggle.com/svencelin/covid19-tweets-afternoon-31032020 Online; accessed 2020.

[58] Echen, COVID-19-tweetids-GIT, 2020, https://github.com/echen102/covid-19-TweetIDs Online; accessed 2020.

[59] U. Qazi, M. Imran, F. Ofli, GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information, SIGSPATIAL Special 12 (1) (2020) 6–15.

[60] N. Reimers, I. Gurevych, 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 (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992, http://dx.doi.org/10.18653/v1/D19-1410, https://www.aclweb.org/anthology/D19-1410.

[61] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[62] S.R. Bowman, G. Angeli, C. Potts, C.D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 632–642, http://dx.doi.org/10.18653/v1/D15-1075, https://www.aclweb.org/anthology/D15-1075.

[63] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, A.E. Hassanien, Sentiment analysis of COVID-19 tweets by deep learning classifiers—A study to show how popularity is affecting accuracy in social media, Appl. Soft Comput. 97 (2020) 106754, http://dx.doi.org/10.1016/j.asoc.2020.106754, http://www.sciencedirect.com/science/article/pii/S156849462030692X.

[64] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, Affective computing and sentiment analysis, in: A Practical Guide to Sentiment Analysis, Springer, 2017, pp. 1–10.

[65] M.S. Akhtar, A. Ekbal, E. Cambria, How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes], IEEE Comput. Intell. Mag. 15 (1) (2020) 64–75.

[66] M.E. Basiri, S. Nemati, M. Abdar, E. Cambria, U.R. Acharya, ABCDM: An attention-based bidirectional CNN-rnn deep model for sentiment analysis, Future Gener. Comput. Syst. 115 (2021) 279–294.

[67] Y. Jang, C.-H. Park, Y.-S. Seo, Fake news analysis modeling using quote retweet, Electronics 8 (12) (2019) 1377.

[68] F. Monti, F. Frasca, D. Eynard, D. Mannion, M.M. Bronstein, Fake news detection on social media using geometric deep learning, 2019, arXiv preprint arXiv:1902.06673.

[69] S. Yardi, D. Boyd, Dynamic debates: An analysis of group polarization over time on twitter, Bull. Sci. Technol. Soc. 30 (5) (2010) 316–327.

[70] U. Arora, W.S. Paka, T. Chakraborty, Multitask learning for blackmarket tweet detection, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 127–130.

[71] S. Singhania, N. Fernandez, S. Rao, 3han: A deep neural network for fake news detection, in: International Conference on Neural Information Processing, Springer, 2017, pp. 572–581.

[72] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[73] N. Ruchansky, S. Seo, Y. Liu, Csi: A hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 797–806.

[74] L. Cui, K. Shu, S. Wang, D. Lee, H. Liu, dEFEND: A system for explainable fake news detection, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2961–2964.