On the Role of Images for Analyzing Claims in Social Media

Gullal S. Cheema[1][0000−0003−4354−9629], Sherzod Hakimov[1][0000−0002−7421−6213], Eric Müller-Budack[1][0000−0002−6802−1241], and Ralph Ewerth[1,2][0000−0003−0918−6297]

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[2] L3S Research Center, Leibniz University Hannover, Germany
{gullal.cheema,sherzod.hakimov,eric.mueller,ralph.ewerth}@tib.eu

Abstract. Fake news is a severe problem in social media. In this paper, we present an empirical study on visual, textual, and multimodal models for the tasks of claim, claim check-worthiness, and conspiracy detection, all of which are related to fake news detection. Recent work suggests that images are more influential than text and often appear alongside fake text. To this end, several multimodal models have been proposed in recent years that use images along with text to detect fake news on social media sites like *Twitter*. However, the role of images is not well understood for claim detection, specifically using transformer-based textual and multimodal models. We investigate state-of-the-art models for images, text (Transformer-based), and multimodal information for four different datasets across two languages to understand the role of images in the task of claim and conspiracy detection.

Keywords: Fake News Detection · Claim Detection · Conspiracy Detection · Multimodal Analysis · Multilingual NLP · Computer Vision · Transformers · COVID-19 · 5G · Twitter

1 Introduction

Social media platforms have become an integral part of our everyday lives, where we use them to connect with people and consume news, entertainment, and buy or sell products. In the last decade, social media has seen exponential growth, with more than a couple of billion users and the increasing presence of prominent people like politicians and celebrities (also called Influencers), organizations, and political parties. On the one hand, this allows influential people or organizations to reach millions of users directly, but it also allows for fake and unverified information to rise and spread faster [42] due to the nature of social media. To deal with misinformation and false claims on online platforms, several independent fact-checking projects like *Snopes, Alt News, Our.News* have been launched that

manually fact-check news and publish their outcomes for public use. Although more such initiatives are coming up worldwide, they cannot keep up with the rate of news or information production on online platforms. Therefore, fake news detection has gathered much interest in computer science for developing automated methods to speed and scale up to handle the continuous fast streaming social media data.

As social media is inherently multimodal in nature, fact-checking initiatives and computation methods consider not only text but also image content [14,21,39,43] as it can be easily fabricated and manipulated due to the availability of free image and video editing tools. In this paper, we investigate the role of images in the context of claim and conspiracy detection. Claim detection is one of the first vital steps to identify fake news where the purpose is to flag a statement if it contains check-worthy facts and information, while the claim may be true or false. Whereas in conspiracy detection, a statement that includes a conspiracy theory is fake news and consists of manipulated facts. Although fake news on social media has been explored recently from a multimodal perspective, images have hardly been considered for claim detection except in recent work by Zlatkova *et al.* [48]. Here, meta-information of images is treated as features, and reverse image search is performed to compare the claim text. However, the image's semantic information is not considered, and the authors highlight that images are more influential than text and appear alongside fake text or unverified news.

Since we are interested in the impact of using images in a multimodal framework, to keep our models simple, we focus on extracting only semantic or contextual features from text and do not consider its structure or syntactic information. To this end, we mainly consider deep transformer Bidirectional Encoder Representations from Transformers (BERT) to extract contextual embeddings and use them along with image embeddings. Taking inspiration from recent work by Cao *et al.* [4], we extract image sentiment features that are widely applied for image credibility or fake news detection in addition to object and scene information for the semantic overlap with textual information.

To carry out this study[3], we experiment with four *Twitter* datasets[4] on binary classification tasks, two of which are from the recent *CLEF-CheckThat! 2020* [2], one in English [36] and the other one in Arabic [17]. The third one is an English dataset from *MediaEval 2020* [33] on conspiracy detection, and the last one is a recent claim detection dataset (English) from Gupta *et al.* [16] on *COVID-19* tweets. Four examples for claim and conspiracy detection are shown in Figure 1. To train our unimodal and multimodal models, we use Support Vector Machines (SVM) [40] and Principal Component Analysis (PCA) [45] for dimensionality reduction due to the small datasets and large size of combined features. We also fine-tune BERT models on the text input to see the extent of the unimodal model's performance on limited-sized datasets and use different pre-trained *BERT* models to see the effect of domain gap. Furthermore,

[3] Code: `https://github.com/cleopatra-itn/image_text_claim_detection`
[4] Dataset: `https://zenodo.org/record/4592249`

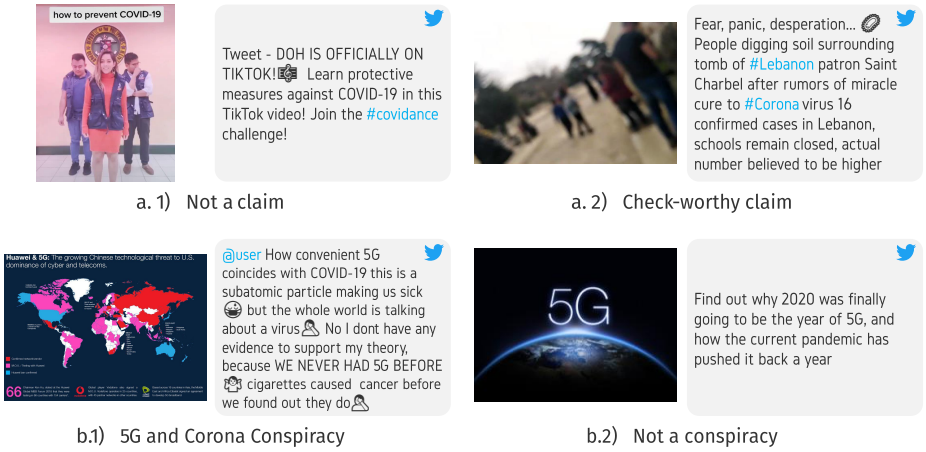

Fig. 1: Examples from CLEF-English [36] (a) check-worthy claims dataset and MediaEval [33] (b) conspiracy detection dataset

we investigate the recently proposed transformer-based *ViLBERT* [25] (Vision-and-Language BERT) model that learns semantic features via co-attention on image and textual inputs. Just like *BERT* models, we perform fixed embedding and fine-tuning experiments using *ViLBERT* to see if a large transformer-based multimodal model can learn meaningful representation and perform better on small-sized datasets.

The remainder of the paper is organized as follows. Section 2 briefly discusses related work on fake news detection and the sub-problems of claim and conspiracy detection. Section 3 presents details of image, text, and multimodal features as well as the fine-tuned and applied models. Section 4 describes the experimental setup, results and summarizes our findings. Section 5 concludes the paper with future research directions.

2 Related Work

There is a wide body of work on fake news detection that goes well beyond this paper's scope. Therefore, we restrict this section to multimodal fake news, claim detection, and conspiracy detection.

2.1 Unimodal Approaches

The earliest claim detection works go back a decade. Rosenthal *et al.* [34] in their pioneering work extracted claims from *Wikipedia* discussion forums. They classified them via logistic regression using the sentiment, syntactic and lexical features like POS (Part-of-Speech) tags and n-grams, and other statistical features

over text. Since then, researchers have proposed context dependent [22], context independent [23], and cross-domain [10] and in-domain approaches for claim detection. Recently, the transformer-based models [6] have replaced structure-based claim detection approaches due to their success in several downstream natural language processing (NLP) tasks.

For claim detection on social media in particular, recently *CLEF-CheckThat! 2020* [2] hosted a challenge to detect check-worthy claims in *COVID-19* related English tweets and several other topics in Arabic. The challenge attracted several models with top submissions [7,32,44] all using some version of transformer-based models like *BERT* [11] and *RoBERTa* [24] along with tweet meta-data and lexical features. Outside of CLEF challenges, some works [12,27] have also conducted a detailed study on detecting check-worthy tweets in U.S. politics and proposed real-time systems to monitor and filter them. Taking inspiration from [10], Gupta *et al.* [16] address the limitations of current methods in cross-domain claim detection by proposing a generalized claim detection model called *LESA* (Linguistic Encapsulation and Semantic Amalgamation). Their model combines contextual transformer features with learnable POS and dependency relation embeddings via transformers to achieve impressive results on several datasets. For conspiracy detection, *MediaEval 2020* [33] saw interesting methods to automatically detect 5G and Coronavirus conspiracy in tweets. Top submissions used BERT [8,28] pre-trained on *COVID* Twitter data, tweet meta-data, graph network data and RoBERTa models [9] along with Graph Convolutional Neural (GCN) networks.

2.2 Multimodal Approaches

For multimodal fake news in general, several benchmark datasets have been proposed in the last few years, generating interest in developing multimodal visual and textual models. In one of the relatively early works, Jin *et al.* [20] explored rumor detection on Twitter using text, social context (emoticons, URLs, hashtags), and the image by learning a joint representation with attention from LSTM outputs over image features. The authors observed the benefit of using the image and social context in addition to text by improving the detection of fake news in Twitter and Weibo datasets. Later, Wang *et al.* [43], proposed an improved model that learns a multi-task model to detect fake news as one task and event discriminator as another task to learn event invariant representations. Since then, improvements have been proposed via using multimodal variational autoencoders [21], transfer learning [15,39] with transformer-based text and deep visual CNN models. Recently, Nakamura [30] *et al.* proposed a fake news dataset *r/Fakeddit* mined from Reddit with over 1 million samples, which includes text, images, meta-data, and comments data. The data is labeled through distant supervision into 2-way, 3-way, and 6-way classification categories. In addition to our different tasks, another difference with the approaches mentioned above is that the size of the datasets is moderate (several thousand) to large (millions) in comparison to a few hundred or a couple of thousand samples in our four datasets for *claim* and *conspiracy detection*.

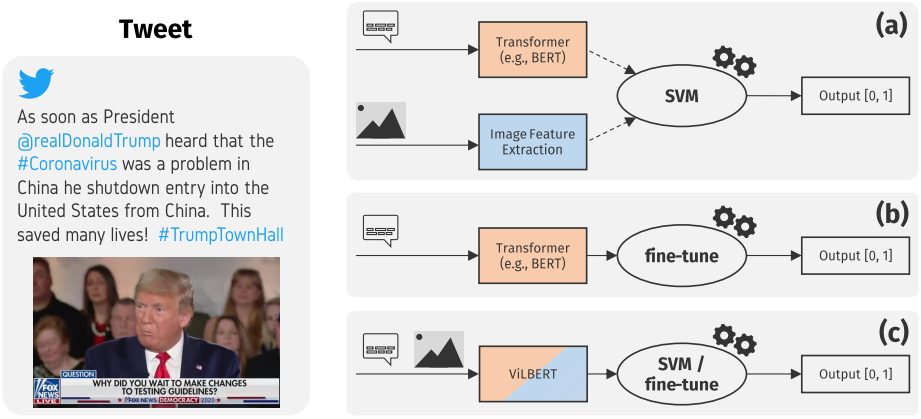

Fig. 2: Workflow of the proposed solutions for claim and conspiracy detection in tweets using multimodal information from text (orange) and image (blue). Three approaches are investigated: (a) Training of an SVM based on the combined features extracted from pre-trained models for visual and/or textual information extraction. (b) Fine-tuning a transformer network (*BERT*) solely using textual information. (c) Fine-tuning *ViLBERT* [25] or training an SVM based on its multimodal embeddings extracted from text and image.

3 Methodology

In this section, we provide details of different image (Section 3.1), textual (Section 3.2), and multimodal (Section 3.3) models and their feature encoding process and how classification models (Section 3.4) are built. An overview of classification models are presented in Figure 2.

3.1 Image Models (I)

The purpose of image models is to encode the presence of different objects, scene, place or background, and affective image content. When learning a multimodal model or a classifier, specific overlapping patterns between image and text can act as discriminatory features for claim detection.

Object Features (I_o) In order to encode objects and the overall image content, we extract features from a pre-trained *ResNet* [19] model trained on *ImageNet* [35] dataset. The pre-trained model has been shown to boost performance over low-level features in several computer vision tasks. We use widely recognized *ResNet*-152 and its last convolution layer to extract features instead of the object categories (final layer). The final convolutional layer outputs 2048 feature maps each of size 7×7, which is then pooled with a global average to get a 2048-dimensional vector.

Place and Scene Features (I_p) In order to encode the scene information in an image, we extract features from a pre-trained *ResNet* [19] model trained on

Places365 [47] dataset. In this case, we use *ResNet*-101 and follow the same encoding process as described for object features.

Hybrid Object and Scene Features (I_h) We also experiment with a hybrid model trained on both *ImageNet* and *Places365* datasets that encodes object and scene information in a single model. To extract these features, we again use a *ResNet*-101 model and follow the same encoding process.

Image Sentiment (I_s) To encode the image sentiment, we use a pre-trained model [41] that is trained on three million images using weak supervision of sentiment label from the tweet text. Although the image labels are noisy, the model has shown superior performance on unseen Twitter testing datasets. We use their best CNN model based on *VGG-19* [38]. The image sentiment embeddings (I_{se}) are extracted from the last layer in the model, which are 4096-dimensional vectors. Additionally, we extract the image sentiment predictions (I_{sp}) from the classification layer that outputs a three-dimensional vector corresponding to the probabilities of three sentiment classes (Negative, Neutral and Positive).

3.2 Textual Models (T)

Since context and semantics of the sentence is shown [2,6] to be important for claim detection, we use transformer-based *BERT*-Base [11] (T_{BB}), to extract contextual word embeddings and employ different pooling strategies to get a single embedding for the tweet. As different layers of *BERT* capture different kinds of information, we experiment with four combinations, i.e., 1) concatenate the last four hidden layers, 2) sum of the last four hidden layers, 3) the last hidden layer, and 4) the second last hidden layer. We finally take an average over the word embeddings to obtain a single vector.

To reduce the domain gap for our Twitter datasets in English, we experiment with two *BERT* models. The first variant is called *BERTweet* [31] (T_{BT}) a BERT-base model that is further pre-trained on 850 million English tweets, and the second one called *COVID-Twitter-BERT* [29] (T_{CT}), a BERT-large model trained on 97 million English tweets on the topic of *COVID-19*. For Arabic tweets, we experiment with the *AraBERT* [1] (T_{AB}) that is trained on Arabic news corpus called *OSIAN* [46] and 1.5 Billion words Arabic corpus [13]. We also perform two experiments, one with raw tweets and the other with pre-processing tweets as part of the *AraBERT's* language-specific text processing method.

For English text, with vanilla BERT-base model, we pre-process the text by following the steps mentioned in Cheema *et. al.* [7] using the publicly available text processing tool Ekphrasis [3]. We also show the performance of vanilla *BERT*-base on raw tweets (T_{BB}^{Raw}) to reflect its sensitivity towards text pre-processing (T_{BB}^{Clean}). For both *BERTweet* and *COVID-Twitter-BERT*, we follow their pre-processing steps, which normalize text, and additionally replaces *user mentions, emails, URLs* with special keywords.

3.3 Multimodal Models (M)

ViLBERT (Vision-and-Language BERT) We use *ViLBERT* [25], one of the recent multimodal transformer architectures that process image and text inputs through two separate transformer-based streams and combines them through transformer layers with the co-attention. It eventually outputs co-attended image and text features that can be combined (added, multiplied or concatenated) to learn a classifier for vision and language tasks. The authors proposed to use visual grounding as a self-supervised pre-training task on a large conceptual captions dataset [37]. They used the model for various downstream tasks involving vision and language, such as visual question answering, visual commonsense reasoning, and caption-based image retrieval.

For the image branch, *ViLBERT* uses state-of-the-art object detection model *Mask R-CNN* [18] and extracts top 100 region proposals (boxes) and their corresponding features. These features are used in a sequence through a 5-layer image transformer, which outputs the image region embeddings. For the text branch, it uses BERT-base model to get the contextual word embeddings. A 6-layer transformer block with the co-attention follows the individual streams that outputs the co-attended image and text embeddings.

Feature Extraction In our fixed embedding experiments with a SVM, we experiment with the output of pooling and last layers of image and text branches. With pooling layers, we directly concatenate (M_{pool}^{CAT}) the image and text outputs. With last layer outputs we average the image region embeddings and word embeddings to get one single embedding per modality and then concatenate them (M_{avg}^{CAT}). From pooling layers, each modality's embedding size is a 1024-dimensional vector, and the last layer average of embeddings gives 1024 and 768-dimensional vectors for image and text, respectively. For fine-tuning, we follow *ViLBERT*'s downstream task approach, where the pooling layer outputs are either added (M_{pool}^{ADD}) or multiplied (M_{pool}^{MUL}) and passed to a classifier. For Arabic text, we use Google Translate to convert the text into English because all *ViLBERT* models are trained on English text.

ViLBERT is fine-tuned on several downstream tasks which can be relevant for encapsulating image-text relationship for our claim detection problem. Therefore, we experiment with four different pre-trained models, namely, conceptual captions , image retrieval (*Image-Ret*), grounding referring expressions (localize an image region given a natural language reference) (*RefCOCO*), and a multi-task model [26] that is trained on 12 different tasks.

3.4 Classification of Tweets

For our fixed embedding experiments, we train SVM models with each type of image and text embeddings for binary classification of tweets as shown in Figure 2 (a). For fine-tuning textual models (Figure 2 (b)), given that we have relatively small-sized datasets, we only experiment with fine-tuning the last two and four layers of transformer models for each dataset. We concatenate the image

and text features for multimodal fixed embedding experiments and train an SVM model over them for classification.

In the case of *ViLBERT* (Figure 2 (c)), we again train SVM over the extracted pooled image and text outputs for classification. For fine-tuning, we fix the individual transformer branches and experiment with fine-tuning the last two and four co-attention layers to activate the interaction between modalities. It enables us to see the effect of only the attention mechanism that can show the benefit of an image and text in claim detection. We use a simple classifier on top of *ViLBERT* outputs as recommended by the authors of *ViLBERT*, which includes a linear layer for down projecting outputs to 128 dimensions, followed by $ReLU$ (Rectified Linear Unit) non-linear activation function, a normalization layer and finally a binary classification layer. Dropout is used to avoid overfitting, and the fine-tuning is performed by minimizing the cross-entropy loss.

4 Experiments and Results

In this section, we describe all the datasets and their statistics, training details and hyper-parameters, model details, experimental results, and discuss them as obtained by different models mentioned in Section 3.

4.1 Datasets

We selected the following four publicly available *Twitter* datasets with high-quality annotations (which excludes [30], besides its focus on fake news), three of which are on claim detection and one on conspiracy detection. The number of tweets in the original datasets is four to fifteen times more as they were mined for text-based fake news detection. We only selected tweets that have an image.
CLEF-En [36] - Released as a part of *CLEF-CheckThat! 2020* challenge, the purpose is to identify *COVID-19* related tweets that are check-worthy claims vs not check-worthy claims. Only 281 English tweets in the dataset include images, whereas the original dataset included 964 tweets.
CLEF-Ar [17] - Released in the same challenge, the dataset consists of 15 topics related to middle east including *COVID-19* and the purpose is to identify check-worthy claims. It consists of 2571 Arabic tweets and corresponding images.
MediaEval [33] - Released in *MediaEval 2020* workshop [33] challenge on identifying 5G and Coronavirus conspiracy tweets. The original dataset has three classes, 5G and Corona conspiracy, other conspiracies, and no conspiracy. To make the problem consistent with other datasets in this paper, we combine conspiracy classes (Corona and others) and treat it as a binary classification problem. It consists of 1724 tweets and images.
LESA [16] - This is a recently proposed dataset of *COVID-19* related tweets on the problem of claim detection. Here, the problem is identifying whether a tweet is a claim or not, and not the claim check-worthiness as in *CLEF-En*. The original dataset consists of 10 000 tweets in English, out of which only 1395 consists of images.

We applied 5-fold cross-validation to overcome the issue of low number of samples in each dataset. We used the ratio of around 72:10:18 for training, validation, and testing in each data split. Next, we report the experimental results for different model configurations. The reported results are averaged across five splits of each dataset. We report accuracy and weighted-F1 measure to account for label imbalance in all the datasets.

4.2 Setup and Hyper-parameters

SVM hyper-parameters: we perform grid search over PCA energy (%) conservation, regularization parameter C and RBF kernel's *gamma*. The parameter range for *PCA* varies from 100% (original features) to 95% with decrements of 1. The parameter range for C and *gamma* vary between -1 to 1 on a log-scale with 15 steps. For experiments only on the *CLEF-En* dataset, we use the range between -2 to 0 for C and *gamma*, as the number of samples are very low and needs aggressive regularization. We normalize the final embedding so that $l2$ norm of the vector is 1.
Fine-tuning BERT and VilBERT: we use a batch size of 4 for *CLEF-En* and 16 for the other datasets. We train all the models for 6 epochs with a starting learning rate of $5e-5$ and a linear decay. A dropout with ratio 0.2 is applied after the first linear layer in the classifier for regularization during fine-tuning.

Table 1: The classification results on all datasets using the textual and visual features (see Sections 3.1 and 3.2). Models marked with† are fine-tuning results and the rest are SVM-based. The best result for each group (bold), and the best result for each dataset (bold and underlined) are highlighted.

Model	**CLEF-En** [36]		**CLEF-Ar** [17]		**LESA** [16]		**MediaEval** [33]	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
I_o	**0.6748**	**0.6180**	0.6991	0.6770	**0.8223**	**0.7775**	0.7189	**0.6968**
I_p	0.6033	0.5966	0.6961	0.6558	0.8159	0.7687	0.7215	0.6355
I_h	0.6551	0.6108	**0.7052**	**0.6776**	0.8223	0.7744	**0.7260**	0.6581
I_{se}	0.6384	0.6223	0.7073	0.6563	0.8143	0.7516	0.708	0.6822
T_{BB}^{Clean}	0.7501	0.7514	-	-	**0.8279**	**0.8015**	0.8130	0.8026
T_{BB}^{Raw}	0.7459	0.7346	-	-	0.8119	0.7873	0.8298	0.8261
T_{BT}	**0.7656**	**0.7661**	-	-	0.8255	0.8023	0.8272	0.8232
T_{CT}	0.7178	0.7123	-	-	0.8175	0.8045	**0.8479**	**0.8479**
T_{AB}	-	-	**0.8362**	**0.8307**	-	-	-	-
$T_{BB}^{Clean\dagger}$	0.6942	0.6804	-	-	0.8319	0.809	0.8046	0.7952
$T_{BT}^{\dagger}$	**0.7420**	**0.7363**	-	-	**0.8486**^{2}	**0.8303**^{2}	0.8407^2	0.8342^2
$T_{CT}^{\dagger}$	0.7146	0.6784	-	-	0.8303	0.8075	**0.8627**	**0.8604**
$T_{AB}^{\dagger}$	-	-	**0.8431**	**0.8432**	-	-	-	-

4.3 Results

Table 1 and Table 2 show the unimodal and multimodal models' performance for all the four datasets based on type of features and feature combinations respectively.

Unimodal Results - In Table 1, it can be seen that all the visual features perform poorly in comparison to textual features. This is expected as visual information on its own cannot indicate whether a social media post makes a claim unless it has text or it's a video. Among the four types of visual models, *Object* (I_o) and *Hybrid* (I_h) features are slightly better, probably because the place or scene information (lowest F1 for all datasets) on its own is not a useful indicator in images for claim detection. With textual features, *BERT* models that are further pre-trained on tweets ($T_{BT}, T_{BT}^{\dagger}$) and *COVID*-related data ($T_{CT}, T_{CT}^{\dagger}$) perform better in comparison to vanilla *BERT* ($T_{BB}^{Clean}, T_{BB}^{Clean\dagger}$) in at-least three datasets. It suggests that the tweets' structure and the domain gap are better captured and reduced respectively in Twitter corpus pre-trained models. Further, normalizing (T_{BB}^{Clean}) the tweet text delivers better performance than using the raw text (T_{BB}^{Raw}). In SVM training, we observed the sum of the last four layers of BERT to compute the embeddings performs better than the other pooling combinations. It indicates that downstream tasks can benefit from the diverse information in different layers of BERT. Similarly, fine-tuning the last four layers instead of two (marked with[2]) gives better performance across all the datasets with *BERT-base* ($T_{BB}^{Clean\dagger}$), *COVID-Twitter-BERT* ($T_{CT}{}^{\dagger}$) and *AraBERT* ($T_{AB}{}^{\dagger}$).

Multimodal Results - In Table 2, we can see the effect of combining visual features with textual features by using a simple concatenation in SVM and also with multimodal co-attention transformer *ViLBERT*. Although we do not see any benefit of using the image sentiment embeddings (I_{se}) in unimodal models, here instead, we use the image sentiment predictions (I_{sp}) that perform better or equivalent in comparison to other visual features. For instance, in case of *CLEF-Ar*, sentiment predictions I_{sp} with *AraBERT* ($T_{AB}{}^{\dagger}$) gives the best fixed embedding performance. Similarly, combining hybrid features (I_h) with *BERT-base* ($T_{BB}^{Clean\dagger}$) and object features with *COVID-Twitter-BERT* ($T_{CT}{}^{\dagger}$) in case of *LESA* and *MediaEval* improves the metrics by 1% over textual SVM models.

With *ViLBERT*, it is interesting to see that with fixed visual and textual branches, it can capture some information from image and text with co-attention to boost performance in case of *LESA* and *MediaEval*. It is worth mentioning that the best unimodal textual models for English and Arabic are pre-trained models further trained on Twitter and language-specific data corpus. In the case of *ViLBERT*, there is a wider domain gap, and for Arabic, the translation process loses quite a bit of information that results in a drop in performance. Different pooling operations applied for pre-trained *ViLBERT* models show more difference in fixed-embedding SVM experiments where the average pooling (M_{avg}^{CAT}) yields a considerable performance, which we also observed in unimodal SVM experiments. We observed that pre-training tasks (best two reported in Table 2) also matter, where image retrieval (Image-Ret) and language reference grounding

Table 2: Multimodal classification results on all datasets based on combination of textual, visual and multimodal features (see Sections 3.1, 3.2 and 3.3). Models marked with† are fine-tuning results on VilBERT and the rest are SVM-based. *ReCOCO* and *Image-Ret* refers to pre-trained *ViLBERT* models. Layers refers to the number of fine-tuned co-attention layers in VilBERT.

Model	**CLEF-En** [36]		**CLEF-Ar** [17]		**LESA** [16]		**MediaEval** [33]	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
Best Unimodal	0.7656	0.7661	0.8431	0.8432	0.8486	0.8303	0.8627	0.8604
T ->	T_{BB}^{Clean}		T_{AB} [1]		T_{BB}^{Clean}		T_{CT} [29]	
$I_o + T$	0.7219	0.7053	0.8054	0.8053	0.8311	0.7953	**0.8594**	**0.8566**
$I_p + T$	0.7336	0.7296	0.8184	0.8168	0.8223	**0.7955**	0.8472	0.8460
$I_h + T$	0.7259	0.7003	0.8085	0.8060	**0.8335**	0.7907	0.8549	0.8527
$I_{sp} + T$	**0.7557**	**0.7575**	**0.8370**	**0.8319**	0.8271	0.8009	0.8485	0.8483
Pre-trained ->	**RefCOCO**		**Image-Ret**		**RefCOCO**		**Image-Ret**	
M_{pool}^{CAT}	0.6980	0.6941	0.7125	0.6990	0.8175	0.7842	0.7357	0.7142
M_{avg}^{CAT}	**0.7062**	**0.7022**	**0.7454**	**0.7245**	**0.8175**	**0.7910**	**0.7892**	**0.7832**
Layers ->	**2**		**4**		**4**		**4**	
$M_{pool}^{ADD\dagger}$	**0.7339**	**0.7322**	0.7449	0.7214	**0.8446**	**0.8196**	0.7989	0.7820
$M_{pool}^{MUL\dagger}$	0.7336	0.7341	**0.7449**	**0.7389**	0.8446	0.8191	0.7937	0.7772
$M_{avg}^{CAT\dagger}$	0.7182	0.7121	0.7466	0.7415	0.8319	0.8063	**0.8014**	**0.7900**

(RefCOCO) features perform much better for all the datasets. It is explainable since both tasks require capturing complex relationships and linking text to specific image regions in the image, enabling them to perform better for our tasks.

4.4 Discussion of Results

We can summarize the findings of our experiments as follows: **1) Domain-specific languages models should be preferred for downstream tasks such as claim detection or fake news**, where underlying meaning and context of certain words (like COVID) is essential, **2) Multimodality certainly helps as seen with multimodal transformer models**, where activating interaction through co-attention layers between fixed unimodal embeddings improves the performance in two datasets, **3) To further understand underlying multimodal dynamics it might be better to explicitly model multimodal relationships**, for instance, importance of image or correlation between image-text in addition to claim detection, **4) Certain pre-training tasks in *ViLBERT* are better suited for downstream tasks** and need further introspection on larger datasets, and lastly, **5) Visual models need to be better adapted to social media images**, for instance, the models used here are not sufficient for

diagrams or images with large text, which constitute around 30-40% of *LESA* and *MediaEval* datasets.

5 Conclusion

In this paper, we have investigated the role of images and tweet text for two problems related to fake news, claim, and conspiracy detection. For this purpose, we combined several state-of-the-art CNN features for images with BERT features for text. We observed the performance improvement over unimodal models in two out of four *Twitter* datasets over two languages. We also experimented with the recently proposed multimodal co-attention transformer *ViLBERT* and observed a promising performance using both image and text even with relatively small-sized datasets. In future work, we will look into other ways to include external knowledge in domain-independent claim detection models without relying on different domain-specific language models. Second, we plan to investigate multimodal transformers in more detail and analyze if the performance does scale with more data in similar tasks. Finally, to address the limitation of visual models, we will consider models that can deal with text and graphs in images and extract suitable features.

Acknowledgements

This work was funded by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no 812997.

References

1. Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. pp. 9–15. European Language Resource Association (2020)
2. Barrón-Cedeno, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., et al.: Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 215–236. Springer (2020)
3. Baziotis, C., Pelekis, N., Doulkeridis, C.: DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 747–754. Association for Computational Linguistics (2017)
4. Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., Li, J.: Exploring the role of visual content in fake news detection. Disinformation, Misinformation, and Fake News in Social Media pp. 141–161 (2020)
5. Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)

6. Chakrabarty, T., Hidey, C., McKeown, K.: IMHO fine-tuning improves claim detection. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 558–563. Association for Computational Linguistics (2019)
7. Cheema, G.S., Hakimov, S., Ewerth, R.: Check_square at checkthat! 2020 claim detection in social media via fusion of transformer and syntactic features. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece. CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)
8. Cheema, G.S., Hakimov, S., Ewerth, R.: Tib's visual analytics group at mediaeval'20: Detecting fake news on corona virus and 5g conspiracy. MediaEval 2020 Workshop (2020)
9. Claveau, V.: Detecting fake news in tweets from text and propagation graph: Irisa's participation to the fakenews task at mediaeval 2020. In: MediaEval 2020 Workshop (2020)
10. Daxenberger, J., Eger, S., Habernal, I., Stab, C., Gurevych, I.: What is the essence of a claim? cross-domain claim identification. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2055–2066. Association for Computational Linguistics (2017)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019)
12. Dogan, F., et al.: Detecting Real-time Check-worthy Factual Claims in Tweets Related to US Politics. Ph.D. thesis (2015)
13. El-Khair, I.A.: 1.5 billion words arabic corpus. ArXiv **abs/1611.04033** (2016)
14. Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France. pp. 877–880. ACM (2019)
15. Giachanou, A., Zhang, G., Rosso, P.: Multimodal multi-image fake news detection. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). pp. 647–654. IEEE (2020)
16. Gupta, S., Singh, P., Sundriyal, M., Akhtar, M.S., Chakraborty, T.: Lesa: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. arXiv preprint arXiv:2101.11891 (2021)
17. Hasanain, M., Haouari, F., Suwaileh, R., Ali, Z., Hamdan, B., Elsayed, T., Barrón-Cedeño, A., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media. In: Cappellato et al. [5]
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy. pp. 2980–2988. IEEE Computer Society (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA. pp. 770–778. IEEE Computer Society (2016)
20. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 795–816 (2017)

21. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: MVAE: multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference, WWW 2019, San Francisco, CA, USA. pp. 2915–2921. ACM (2019)
22. Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., Slonim, N.: Context dependent claim detection. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1489–1500. Dublin City University and Association for Computational Linguistics (2014)
23. Lippi, M., Torroni, P.: Context-independent claim detection for argument mining. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina. pp. 185–191. AAAI Press (2015)
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
25. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada. pp. 13–23 (2019)
26. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA. pp. 10434–10443. IEEE (2020)
27. Majithia, S., Arslan, F., Lubal, S., Jimenez, D., Arora, P., Caraballo, J., Li, C.: ClaimPortal: Integrated monitoring, searching, checking, and analytics of factual claims on Twitter. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 153–158. Association for Computational Linguistics (2019)
28. Manh Duc Tuan, N., Quang Nhat Minh, P.: Fakenews detection using pre-trained language models and graph convolutional networks. In: MediaEval 2020 Workshop (2020)
29. Müller, M., Salathé, M., Kummervold, P.E.: Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503 (2020)
30. Nakamura, K., Levy, S., Wang, W.Y.: Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6149–6157. European Language Resources Association (2020)
31. Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14. Association for Computational Linguistics (2020)
32. Nikolov, A., Da San Martino, G., Koychev, I., Nakov, P.: Team_Alex at CheckThat! 2020: Identifying check-worthy tweets with transformer models. In: Cappellato et al. [5]
33. Pogorelov, K., Schroeder, D.T., Burchard, L., Moe, J., Brenner, S., Filkukova, P., Langguth, J.: Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020. In: MediaEval 2020 Workshop (2020)
34. Rosenthal, S., McKeown, K.: Detecting opinionated claims in online discussions. In: 2012 IEEE sixth international conference on semantic computing. pp. 30–37. IEEE (2012)

35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
36. Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeño, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Da San Martino, G., Nakov, P.: Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In: Cappellato et al. [5]
37. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565. Association for Computational Linguistics (2018)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, Conference Track Proceedings (2015)
39. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multi-modal framework for fake news detection. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). pp. 39–47. IEEE (2019)
40. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural processing letters **9**(3), 293–300 (1999)
41. Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., Tesconi, M.: Cross-media learning for image sentiment analysis in the wild. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 308–317 (2017)
42. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
43. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: EANN: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK. pp. 849–857. ACM (2018)
44. Williams, E., Rodrigues, P., Novak, V.: Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece. CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)
45. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)
46. Zeroual, I., Goldhahn, D., Eckart, T., Lakhouaja, A.: OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop. pp. 175–182. Association for Computational Linguistics (2019)
47. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
48. Zlatkova, D., Nakov, P., Koychev, I.: Fact-checking meets fauxtography: Verifying claims about images. 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). pp. 2099–2108. Association for Computational Linguistics (2019)