MM-Claims: A Dataset for Multimodal Claim Detection in Social Media

Gullal S. Cheema[1], Sherzod Hakimov[1,3], Abdul Sittar[2], Eric Müller-Budack[1,3], Christian Otto[3], and Ralph Ewerth[1,3]

[1]TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]Jozef Stefan Institute, Ljubljana, Slovenia
[3]L3S Research Center, Leibniz University Hannover, Germany
{gullal.cheema, sherzod.hakimov, eric.mueller}@tib.eu
{christian.otto, ralph.ewerth}@tib.eu
abdul.sittar@ijs.si

Abstract

In recent years, the problem of misinformation on the web has become widespread across languages, countries, and various social media platforms. Although there has been much work on automated fake news detection, the role of images and their variety are not well explored. In this paper, we investigate the roles of image and text at an earlier stage of the fake news detection pipeline, called claim detection. For this purpose, we introduce a novel dataset, *MM-Claims*, which consists of tweets and corresponding images over three topics: *COVID-19*, *Climate Change* and broadly *Technology*. The dataset contains roughly 86 000 tweets, out of which 3400 are labeled manually by multiple annotators for the training and evaluation of multimodal models. We describe the dataset in detail, evaluate strong unimodal and multimodal baselines, and analyze the potential and drawbacks of current models.

1 Introduction

The importance of combating misinformation was once again illustrated by the coronavirus pandemic, which came along with a lot of "potentially lethal" misinformation. At the beginning of the COVID-19 pandemic, the United Nations (UN) (DGC, 2020) started even using the term “infodemic” for this phenomenon of misinformation and called for proper dissemination of reliable facts. However, tackling misinformation online and specifically on social media platforms is challenging due to the variety of information, volume, and speed of streaming data. As a consequence, several studies have explored different aspects of COVID-19 misinformation online including sharing patterns (Pennycook et al., 2020), platform-dependent engagement patterns (Cinelli et al., 2020), web search behaviors (Rovetta and Bhagavathula, 2020), and fake images (Sánchez and Pascual, 2020).

We are primarily interested in claims on social media from a multimodal perspective (Figure 1).

Figure 1: Examples for each of the four classes in the MM-Claims dataset: a) **not a claim** (both image and text together abstractly represent effects of climate change), b) **claim but not checkworthy** (claim in text, but lacks details like to which experts is referred to, while image is relevant), c) **checkworthy** but not visually relevant (claim in text targets CDC and China but the image is a stock photograph), and d) **checkworthy and visually relevant** (claim in text and in image with important details in both).

Claim detection can be seen as an initial step in fighting misinformation and as a precursor to prioritize potentially false information for fact-checking. Traditionally, claim detection is studied from a linguistic standpoint where both syntax (Rosenthal and McKeown, 2012) and semantics (Levy et al., 2014) of the language matter to detect a claim accurately. However, claims or fake news on social media are not bound to just one modality and become a complex problem with additional modalities like images and videos. While it is clear that a claim in the text is denoted in verbal form, it can also be part of the visual content or as overlaid text in the

Findings of the Association for Computational Linguistics: NAACL 2022, pages 962 - 979
July 10-15, 2022 ©2022 Association for Computational Linguistics

image. Even though much effort has been spent on the curation of datasets (Boididou et al., 2016; Nakamura et al., 2020; Jindal et al., 2020) and the development of computational models for multimodal fake news detection on social media (Ajao et al., 2018; Wang et al., 2018; Khattar et al., 2019; Singhal et al., 2019), hardly any research has focused on multimodal claims (Zlatkova et al., 2019; Cheema et al., 2020b).

In this paper, we extend the definitions of claims and check-worthiness from previous work (Barrón-Cedeno et al., 2020; Nakov et al., 2021) to multimodal claim detection and introduce a novel dataset called *Multimodal Claims (MM-Claims)* curated from Twitter to tackle this critical problem. While previous work has focused on factually-verifiable check-worthy (Barrón-Cedeno et al., 2020; Alam et al., 2020) or general claims (i.e., not necessarily factually-verifiable, e.g., (Gupta et al., 2021)) on a single topic, we focus on three different topics, namely *COVID-19*, *Climate Change* and *Technology*. As shown in Figure 1, MM-Claims aims to differentiate between tweets without claims (Figure 1a) as well as tweets with claims of different types: *claim but not check-worthy* (Figure 1b), *check-worthy claim* (Figure 1c), and *check-worthy visually relevant claim* (Figure 1d). Our contributions can be summarized as follows:

- a novel dataset for multimodal claim detection in social media with more than 3000 manually annotated and roughly 82 000 unlabeled image-text tweets is introduced;
- we present details about the dataset and the annotation process, class definitions, dataset characteristics, and inter-coder agreement;
- we provide a detailed experimental evaluation of strong unimodal and multimodal models highlighting the difficulty of the task as well as the role of image and text content.

The remainder of the paper is structured as follows. Section 2 describes the related work on unimodal and multimodal approaches for claim detection. The proposed dataset and the annotation guidelines are presented in Section 3. We discuss the experimental results of the compared models in Section 4, while Section 5 concludes the paper and outlines areas of future work.

2 Related Work

2.1 Text-based Approaches

Before research on claim detection targeted social media, pioneering work by Rosenthal and McKeown (2012) focused on claims in *Wikipedia* discussion forums. They used lexical and syntactic features in addition to sentiment and other statistical features over text. Since then, researchers have proposed context-dependent (Levy et al., 2014), context-independent (Lippi and Torroni, 2015), cross-domain (Daxenberger et al., 2017), and in-domain approaches for claim detection. Recently, transformer-based models (Chakrabarty et al., 2019) have replaced structure-based claim detection approaches due to their success in several Natural Language Processing (NLP) downstream tasks. A series of workshops (Barrón-Cedeno et al., 2020; Nakov et al., 2021) focused on claim detection and verification on Twitter and organized challenges with several sub-tasks on text-based claim detection around the topic of *COVID-19* in multiple languages. Gupta et al. (2021) addressed the limitations of current methods in cross-domain claim detection by proposing a new dataset of about ∼10 000 claims on *COVID-19*. They also proposed a model that combines transformer features with learnable syntactic feature embeddings. Another dataset introduced by Iskender et al. (2021) includes tweets in German about *climate change* for claim and evidence detection. Wührl and Klinger (2021) created a dataset for biomedical Twitter claims related to *COVID-19, measles, cystic fibrosis and depression*. One common theme and challenge among all the datasets is the variety of claims where some types of claims (like implicit) are harder to detect than explicit ones where a typical claim structure is present. Table 1 shows a comparison of existing social media based claim datasets, with number of samples, modalities, data sources, language, topic, and type of tasks.

2.2 Multimodal Approaches

From the multimodal perspective, very few works have analyzed the role of images in the context of claims. Zlatkova et al. (2019) introduced a dataset that consists of claims and is created from the idea of investigating questionable or outright false images which supplement fake news or claims. The authors used reverse image search and several image metadata features such as tags from Google Vision API, URL domains and categories, relia-

Datasets	#Samples	Modality	Data source	Language	Topic	Task(s)
Zlatkova et al. (2019)*	1233	Image, Text	Snopes, Reuters	English	Multi-topic	True vs False
Nakov et al. (2021)	18, 014†	Text	Twitter	Multi	Multi-topic†	check-worthiness estimation
Gupta et al. (2021)	9981	Text	Twitter	English	COVID-19	claim detection
Iskender et al. (2021)	300 pairs	Text	Twitter	German	Climate change	claim, evidence detection
Wührl and Klinger (2021)	1200	Text	Twitter	English	Biomedical & COVID-19	claim & claim type detection
MM-Claims (Ours)	3400	Image, Text	Twitter	English	COVID-19, Climate Change, Technology	claim, check-worthiness, visual relevance

Table 1: Comparison of social media based claim datasets. *Zlatkova et al. (2019) is a mix of actual news photographs (from Reuters) and possibly fake images (from Snopes), which went viral on social media sites like Reddit. † 1312 samples are in English and only on the topic of COVID-19.

bility of the image source, etc. Similarly, Wang et al. (2020) performed a large-scale study by analyzing manipulated or misleading images in news discussions on forums like *Reddit*, *4chan* and *Twitter*. For claim detection, Cheema et al. (2021) extended the text-based claim detection datasets of Barrón-Cedeno et al. (2020) and Gupta et al. (2021) with images to evaluate multimodal detection approaches. Although previous work has provided multimodal datasets on claims, they are either on veracity (true or false) of claims or labeled only text-based for a single topic (COVID-19). In terms of multimodal models for image-text data, most previous work is in the related area of multimodal fake news, where several benchmark datasets and models exist for fake news detection (Nakamura et al., 2020; Boididou et al., 2016; Jindal et al., 2020) . In an early work, Jin et al. (2017) explored rumor detection on Twitter using text, social context (emoticons, URLs, hashtags), and the image by learning a joint representation in a deep recurrent neural network. Since then, several improvements have been proposed, such as multi-task learning with an event discriminator (Wang et al., 2018), multimodal variational autoencoder (Khattar et al., 2019) and multimodal transfer learning using transformers for text and image (Giachanou et al., 2020; Singhal et al., 2019).

3 MM-Claims Tasks and Dataset

This section describes the problem of multimodal claim detection (Section 3.1), the data collection (Section 3.2), the guidelines for annotating multimodal claims (Section 3.3), and the annotation process (Section 3.4) to obtain the new dataset.

3.1 Task Description

Given a tweet with a corresponding image, the task is to identify important factually-verifiable or check-worthy claims. In contrast to related work, we introduce a novel dataset for claim detection that is labeled based on both the tweet and the corresponding image, making the task truly multimodal. Our scope of claims is motivated by Alam et al. (2020) and Gupta et al. (2021), which have provided detailed annotation guidelines. We restrict our dataset to factually-verifiable claims (as in Alam et al. (2020)) since these are often the claims that need to be prioritized for fact-checking or verification to limit the spread of misinformation. On the other hand, we also include claims that are personal opinions, comments, or claims existing at sub-sentence or sub-clause level (as in Gupta et al. (2021)), with the condition that they are factually-verifiable. Subsequently, we extend the definition of claims to images along with factually-verifiable and check-worthy claims.

3.2 Data Collection

In previous work on claim detection in tweets, most of the publicly available English language datasets (Alam et al., 2020; Barrón-Cedeno et al., 2020; Gupta et al., 2021; Nakov et al., 2021) are text-based and on a single topic such as *COVID-19*, or *U.S. 2016 Elections*. To make the problem interesting and broader, we have collected tweets on three topics, *COVID-19*, *Climate Change* and broadly *Technology*, that might be of interest to a wider research community. Next, we describe the steps for crawling and preprocessing the data.

3.2.1 Data Crawling

We have used an existing collection of tweet IDs, where some are topic-specific Twitter dumps, and extracted tweet text and the corresponding image to create a novel multimodal dataset.

COVID-19: We combined tweets from three Twitter resources (Banda et al., 2020; Dimitrov et al., 2020; Lamsal, 2020) that were posted between October 2019 and April 2020. In our dataset, we use tweets in the period from March - April 2020.

Climate Change: We used a Twitter resource (Littman and Wrubel, 2019) that contains tweet IDs related to climate change from September 2017 to May 2019. The tweets were originally crawled based on hashtags like *climatechange, climatechangeisreal, actonclimate, globalwarming, climatedeniers, climatechangeisfalse*, etc.

Technology: For the broad topic of *Technology*, we used the *TweetsKB* (Fafalios et al., 2018) corpus. To avoid the extraction of all the tweets from 2019 to 2020 irrespective of the topic, we followed a two-step process to find tweets remotely related to technology. The corpus is available in form of RDF (Resource Description Framework) triples with attributes like tweet ID, hashtags, entities and emotion labels, but without tweet text or media content details. First, we selected tweet IDs based on hashtags and entities, and only kept those that contain keywords like *technology, cryptocurrency, cybersecurity, machine learning, nano technology, artificial intelligence, IOT, 5G, robotics, blockchain*, etc. The second step of filtering tweets based on a selected set of hashtags for each topic is described in the next subsection.

From the above resources, we collected 214 715, 28 374 and 417 403 tweets for the topics *COVID-19*, *Climate Change* and *Technology*, respectively.

3.2.2 Data Filtering

We perform a number of filtering steps to remove inconsistent samples: 1) tweets that are not in English or without any text, 2) duplicated tweets based on tweet IDs, processed text and retweets, 3) tweets with corrupted or no images, 4) tweets with images of less than 200×200 pixels resolution, 5) tweets that have more than six hashtags, and finally, 6) we make a list of the top 300 hashtags in each topic based on count and manually select those related to the selected topics. We only keep those tweets where all hashtags are in the list of top selected hashtags. The hashtags are manually marked because some top hashtags are not relevant to the main topic of interest. The statistics of tweets after each filtering step are provided in the Appendix (see Table 8). In summary, we end up with 17 771, 4874, and 62 887 tweets with images for *COVID-19*, *Climate Change* and *Technology*, respectively.

3.3 Annotation Guidelines

In this section, we provide definitions for all investigated claim aspects, the questions asked to annotators, and the cues and explanations for the annotation questions. We define a claim as *state or assert that something is the case, typically without providing evidence or proof* using the definition in the Oxford dictionary (like Gupta et al. (2021)).

The definition of a *factually-verifiable claim* is restricted to claims that can possibly be verified using external sources. These external sources can be reliable websites, books, scientific reports, scientific publications, credible fact-checked news reports, reports from credible organizations like World Health Organization or United Nations. Although we did not provide external links of reliable sources for the content in the tweet, we highlighted named entities that pop-up with the text and image description. External sources are not important at this stage because we are only interested in marking claims, which have possibly incorrect details and information. A list of identifiable cues (extended from Barrón-Cedeno et al. (2020)) for factually-verifiable claims is provided in the Appendix A.3.1.

To define check-worthiness, we follow Barrón-Cedeno et al. (2020) and identify claims as check-worthy if the information in the tweet is, 1) *harmful* (attacks a person, organization, country, group, race, community, etc), or 2) *urgent or breaking news* (news-like statements about prominent people, organizations, countries and events), or 3) *up-to-date* (referring to recent official document with facts, definitions and figures). A detailed description of these cases is provided in the Appendix A.3.1. Given these key points, the answer to whether the claim is check-worthy is subjective since it depends on the person's (annotator's) background and knowledge.

Annotation Questions: Based on the definitions above, we decided on the following annotation questions in order to identify factually-verifiable claims in multimodal data.

- Q1: *Does the image-text pair contain a factually-verifiable claim? - Yes / No*

- Q2: *If "Yes" to Q1, Does the claim contain harmful, up-to-date, urgent or breaking-news information? - Yes / No*
- Q3: *If "Yes" to Q1, Does the image contain information about the claim or the claim itself (in the overlaid text)? - Yes / No*

Question 3 (Q3) intends to identify whether the visual content contributes to a tweet having factually-verifiable claims. The question is answered "Yes" if one of the following cases hold true: 1) there exists a piece of evidence (e.g. an event, action, situation or a person's identity) or illustration of certain aspects in the claim text, or 2) the image contains overlay text that itself contains a claim in a text form. Please note that we asked the annotators to label tweets with respect to the time they were posted. During our annotation dry runs we observed that there were several false annotations for the tweets where the claims were false but already well known facts. This aspect intends to ignore the veracity of claims since some of the claims become facts over time. In addition, we ignore tweets that are questions and label them as not claims unless the corresponding image consists of a response to the question and is a factually-verifiable claim.

3.4 Annotation Process

Each annotator was asked to answer these questions by looking at both image and text in a given tweet. We distribute the data among nine external and four expert internal annotators for the annotation of training and evaluation splits, respectively. The nine annotators are graduate students with engineering or linguistics background. These annotators were paid 10 Euro per hour for their participation. The four expert annotators are doctoral and postdoctoral researchers of our group with a research focus on computer vision and multimodal analytics. Each annotator was shown a tweet text with its corresponding image and asked to answer the questions presented in Section 3.3. Exactly three annotators labeled each sample, and we used a majority vote to obtain the final label.

3.4.1 Claim Categories

We selected a total of 3400 tweets for manual annotation of training (annotated by external annotators) and evaluation (annotated by internal experts) splits. Each split contains an equal number of samples for the topics: *COVID-19*, *Climate Change*, and *Technology*. Labels for three types of claim[1] annotations are derived:

- binary claim classes: *not a claim*, and *claim*
- tertiary claim classes: *not a claim*, *claim but not check-worthy*, and *check-worthy claim*
- visual claim classes: *not a claim*, *visually-irrelevant claim*, and *visually-relevant claim*

3.4.2 Annotator Training

The annotators were trained with detailed annotation guidelines, which included the definitions given in Section 3.3 and multiple examples. To ensure the quality, we performed two dry runs using a set of samples (30-40) to annotate. Afterwards, the annotations were discussed to check agreements among annotators and the guidelines were refined based on the feedback.

3.4.3 Inter-Annotator Agreement

We measured the agreements between two groups of annotators using *Krippendorff's alpha* (Krippendorff, 2011). The agreements were computed for the three types of annotations described in the previous section. For the training dataset group, we observe 0.53, 0.39, and 0.42 as agreement scores for the *binary*, *tertiary*, and *visual claims*, respectively. For the test dataset group, we observe the following agreement scores: 0.57, 0.47, and 0.52 for three classifications, respectively. The moderate agreement scores suggest that the problem of identifying check-worthy claims is partially a subjective task for both non-experts and experts.

3.4.4 Conflict Resolution Strategy

While a majority is always possible for the binary claim classification that allows us to derive unambiguous labels, entirely different labels could be chosen for the tertiary and visually-relevant claim classification task since the annotators assign three possible classes. Consequently, it is not possible to derive a label with majority voting when each annotator selects a different option. In such cases, we resolve the conflict by prioritizing the *claim but not check-worthy* class since check-worthiness is a stricter constraint and chosen by only one annotator, while two annotators agreed it is a claim. First row in Table 2 shows this case when two annotators indicated that the given sample is a claim (A-2 $\rightarrow$ Q1-Yes, A-3 $\rightarrow$ Q1-Yes). For visual claims, we select a *visually-relevant claim* since it is possible

[1]Here claim is a factually-verifiable claim not any claim

A-1	A-2	A-3	Derived Class
Q1-No	Q1-Yes Q2-No	Q1-Yes Q2-Yes	*Claim but not check-worthy*
Q1-No	Q1-Yes Q3-No	Q1-Yes Q3-Yes	*Visually relevant claim*

Table 2: Conflict resolution strategies to derive class labels where a majority vote can not be reached among three annotators (A) for check-worthiness and visual relevance tasks.

that image and text are related, even when one annotator marked "no" to the claim question. See row two in the table, where one annotator marked "no" to the claim question (A-1 $\rightarrow$ Q1-No), but at least one annotator indicated that the sample is a visually-relevant claim (A-3 $\rightarrow$ Q3-Yes).

3.5 The MM-Claims Dataset

As a result of the annotation process, the *Multimodal Claims (MM-Claims)* dataset[2] consists of 2815 (T_C (training)) and 585 (E_C (evaluation)) samples (C in the subscript stands for "with resolved conflicts"). However, as discussed above, there are conflicting examples for the tertiary and visual claim labels. To train and evaluate our models on unambiguous examples, we derive a subset of *Multimodal Claim (MM-Claims)* dataset that contains 2555 (T) and 525 (E) samples "without conflicts" where a majority vote can be taken. We divided the training set (T_C, T) in each case further into training and validation in a 90:10 split for hyper-parameter tuning.

We noticed that one-third of the images in the dataset contains a considerable amount of overlaid text (five or more words). As suggested by previous work (Cheema et al., 2021; Parcalabescu et al., 2021; Kirk et al., 2021), overlaid text in images should be considered in addition to tweet text and other image content. Specifically, the images with overlaid text not only act as related information to the tweet text but are sometimes the central message of the tweet. We used Tesseract-OCR (Fayez, 2021) to select images that contain five or more words in their overlay text. In an internal pre-test with 100 images, we observed that Tesseract-OCR produced more random (and incorrect) text from images than Google Vision API. To reduce the incorrect text, we ran Google Vision API on the selected images (avoiding unnecessary costs) in the second step that resulted in a better quality OCR detected text. Besides the labeled dataset, we will also provide the images, tweet text, and the overlay text (extracted using OCR methods as described above) of the unlabeled portion of the dataset.

[2]Source code is available at: `https://github.com/TIBHannover/MM_Claims`
Dataset (Tweet IDs) and labels are available at: `https://data.uni-hannover.de/dataset/mm_claims`
For complete labeled data access (Images and Tweets), please contact at *gullal.cheema@tib.eu* or *gullalcheema@gmail.com*

4 Experimental Setup and Evaluation

In this section, we describe the features, baseline models, and the comprehensive experiments using our novel dataset. We test a variety of features and recent multimodal state-of-the-art models.

4.1 Features

Pre-processing: For images, we use the standard pre-processing of resizing and normalizing an image, whereas text is cleaned and normalized according to Cheema et al. (2020a) using the Ekphrasis (Baziotis et al., 2017) tool. Besides digits and alphabets, we also keep punctuation to reflect the syntax and style of a written claim.
Image Features: For image encoding, we use a *ResNet-152* (He et al., 2016) model trained on *ImageNet* (Russakovsky et al., 2015) and extract the 2048-dimensional feature vector from the last pooling layer.
Text Features: For encoding tweet and OCR text, we test *BERT* (Devlin et al., 2019) uncased models to extract contextual word embeddings. For classification using Support Vector Machine (SVM, (Cortes and Vapnik, 1995)), we employ a pooling strategy by adding the last four layers' outputs and then average them to obtain the final 768-dimensional vector.
Multimodal Features: We use the following two pre-trained image-text representation learning architectures to extract multimodal features.
The ***ALBEF*** (ALign BEfore Fuse) embedding (Li et al., 2021) results from a recent multimodal state-of-the-art model for vision-language downstream tasks. It is trained on a combination of several image captioning datasets ($\sim$14 million image-text pairs) and uses *BERT* and a visual transformer (Dosovitskiy et al., 2021) for text and image encoding, respectively. It produces a multimodal embedding of 768 dimensions.
The ***CLIP*** (Contrastive Language-Image Pretraining) model (Radford et al., 2021) is trained without any supervision on 400 million image-text pairs.

We evaluate several image encoder backbones including *ResNet* and vision transformer (Dosovitskiy et al., 2021). The *CLIP* model outputs two embeddings of same size, i.e., the image ($CLIP_I$) and the text ($CLIP_T$) embedding, while $CLIP_{I \oplus T}$ denotes the concatenation of two embeddings.

4.2 Training Baselines

In the following, we describe training details, hyper-parameters, input combinations, and different baseline models' details.

4.2.1 SVM

To obtain unimodal and multimodal embeddings for our experiments, we first use PCA (Principal Component Analysis) to reduce the feature size and train a SVM model with the *RBF* kernel. 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 multimodal experiments, image and text embeddings are concatenated before passing them to PCA and SVM. We normalize the final embedding so that *l2* norm of the vector is 1.

4.2.2 BERT and ALBEF Fine-tuning (FT)

We experiment with fine-tuning the last few layers of unimodal and multimodal transformer models to get a strong multimodal baseline and see whether introducing cross-modal interactions improves claim detection performance. We fine-tune the last layers of both the models and report the best ones in Table 3. Additional experimental results on fine-tuned layers are provided in Appendix A.2.5. For fine-tuning, we limit the tweet text to the maximum number of tokens (91) seen in a tweet in the training data and pad the shorter tweets with zeros. Hyper-parameter details for fine-tuning are provided in the Appendix A.1.

4.2.3 Models with OCR Text

To incorporate OCR text embeddings into our models, we experiment with two strategies for embedding generation and one strategy to fine-tune models. To obtain an embedding for SVM models, we experimented with concatenating the OCR embedding to image and tweet text embeddings as well as adding the OCR embedding directly to tweet text embedding. To fine-tune the models, we concatenate the OCR text to tweet text and limit the OCR text to 128 tokens.

4.2.4 State-of-the-Art Baselines

We compare our models with two state-of-the-art approaches for multimodal fake news detection.

MVAE (Khattar et al., 2019) is a multimodal variational auto-encoder model that uses a multi-task loss to minimize the reconstruction error of individual modalities and task-specific cross-entropy loss for classification. We use the publicly available source code and hyper-parameters for our task.

SpotFake (Singhal et al., 2019) is a model built as a shallow multimodal neural network on top of *VGG-19* image and *BERT* text embeddings using a cross-entropy loss. We re-implement the model in PyTorch and use the hyper-parameter settings given in the paper.

4.3 Results

We report accuracy (Acc) and Macro-F1 (F1) for binary (BCD) and tertiary claim detection (TCD) in Table 3. We also present the fraction (in %) of visually-relevant and visually-irrelevant (textual only) claims retrieved by each model in Table 4. Please note that in Table 3 and Table 6, BCD results are shown for only one split ($T_C \rightarrow E_C$), because there are no conflicts in the labels for binary claim classification. Although we do not train the models specifically to detect visual claim labels, we analyze the fraction of retrieved samples in order to evaluate the bias of binary classification models towards a modality.

4.3.1 Impact of Annotation Disagreements

As mentioned in Section 3, we observed disagreements in the annotated data that reflect the real-world difficulty and subjectivity of the problem. Therefore, we analyze the effect of keeping (T_C, E_C) and removing (T, E) conflicting examples in training and evaluation data splits (Table 3, 6). The findings are as follows: 1) multimodal models are more sensitive to the conflict resolution strategy as most have lower accuracy when trained on T_C but relatively better F1 score. On the contrary, visual and textual models perform better on both metrics with training on T_C, 2) overall, training on T_C with conflict resolution is a better strategy with a higher F1 score, i.e., better on claim and check-worthiness (fewer samples) detection; and 3) when comparing all the cross-split experiments in Table 3 and Table 6, multimodal models perform the best in case

Task →	BCD		TCD			
Data Splits →	$T_C \to E_C$		$T \to E_C$		$T_C \to E_C$	
Models ↓	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Random	50.7	50.2	33.3	30.6	33.3	30.6
Majority	62.7	38.5	56.2	35.9	56.2	35.9
ImageNet	63.1	62.6	58.3	42.9	58.5	43.9
CLIP$_I$	70.0	69.8	64.1	50.5	62.4	48.7
BERT	80.5	79.9	71.9	54.1	69.6	59.8
↪ FT	80.9	80.1	72.5	54.5	**75.4**	**64.6**
CLIP$_T$	75.6	74.7	70.6	53.4	67.4	54.5
BERT ⊕ ImageNet	**81.4**	80.9	72.7	57.6	71.6	56.9
↪ ⊕ OCR	80.9	80.4	72.8	58.2	71.9	58.6
CLIP$_{I \oplus T}$	77.8	77.4	71.6	52.9	68.4	54.6
CLIP$_I$ ⊕ BERT	80.3	79.7	72.7	57.9	69.4	59.7
ALBEF	76.9	76.5	71.5	56.1	65.6	57.3
↪ FT	80.2	79.7	**74.5**	**60.7**	72.5	61.0
↪ ⊕ OCR ⊕ FT	**81.4**	**81.1**	72.7	58.2	73.0	60.8
MVAE	64.1	62.9	60.0	41.2	59.7	44.8
SpotFake	71.8	71.4	67.0	49.5	66.3	52.2

Table 3: Accuracy (Acc) and Macro-F1 (F1) for binary (BCD) and tertiary claim detection (TCD) in percent [%]. As described in Section 3.5, we use the training split (T) with resolved (index C) and without (no index) conflicts, and evaluation (test) split (E_C) with conflicts. This evaluation split reflects the real-world scenario for the subjective task of tertiary claim classification (TCD). Unless FT (fine-tuning) is written, all models (except MVAE and SpotFake) are SVM models trained on extracted features.

of "without conflicts" T and E splits. The latter two observations also apply to retrieval of visually-relevant and visually-irrelevant claims in Table 4 and Table 7.

Data Splits →	$T \to E_C$		$T_C \to E_C$	
Models ↓	**V (111)**	**T (145)**	**V (111)**	**T (145)**
ImageNet	35.1	39.3	61.3	57.9
CLIP$_I$	70.3	67.6	**76.6**	73.8
BERT	49.6	76.6	57.7	82.1
↪ FT	52.3	75.9	55.9	**82.8**
CLIP$_T$	46.9	73.1	54.9	73.1
BERT ⊕ ImageNet	57.7	66.2	71.2	77.9
↪ ⊕ OCR	65.8	75.9	71.2	79.3
CLIP$_{I \oplus T}$	65.8	66.9	72.9	75.2
CLIP$_I$ ⊕ BERT	57.7	72.4	57.7	**82.8**
ALBEF	61.2	75.2	63.9	77.9
↪ FT	62.2	77.2	70.3	78.6
↪ ⊕ OCR ⊕ FT	**71.2**	**79.3**	75.7	82.1

Table 4: Visually-relevant (V) and visually-irrelevant (text-only) (T) claim detection evaluation. The number of test samples is reported in brackets and the fraction, how many of them were retrieved, is given in percent [%]. The underlying models are trained for binary claim detection (BCD). The labels for visual relevance are only used for retrieval evaluation.

4.3.2 Results for Unimodal Models

For image-based models, *CLIP$_I$* performs (70.0, 69.8) considerably better than *ResNet-152*'s *ImageNet* (63.1, 62.6) features in terms of both accuracy and F1 metrics (Table 3, block 2). This result is compliant to previous work (Kirk et al., 2021) where the task has a variety of information and text in images. It is further exaggerated and clearly observable in Table 4 where fraction of visually-relevant claims retrieved using *CLIP$_I$* (70.3) is higher and comparable to fine-tuned *ALBEF* ⊕ OCR (71.2).

For text-based models, fine-tuning (FT) *BERT* gives the best performance, better than any other unimodal model. This result indicates that the problem is inherently a text-dominant task. The model also retrieves the most visually-irrelevant claims when trained on T_C. It should be noted that textual models can still identify visually-relevant claims since they can have a claim or certain cues in the tweet text that refer to the image. Finally, the *CLIP$_T$* features perform considerably worse than *BERT* features, possibly because *CLIP* is limited to short text (75 tokens) and is not trained like vanilla *BERT* on a large text corpus.

4.3.3 Results for Multimodal Models

For multimodal models, the combination of *BERT* and *ResNet-152* features performs slightly better ($0.5-1\%$) on two metrics in Table 3 on full dataset in binary task and with T split training in case of tertiary. Although this gain is not impressive, the benefit of combining two modalities is more obvious in identifying visually-relevant claims ($> 10\%$) in Table 4, which comes at the cost of a lower fraction of visually-irrelevant claims. Similarly with *CLIP*, the combination of image and text features (CLIP$_{I \oplus T}$) improves the overall accuracy from CLIP$_I$ or CLIP$_T$. However, we do not see the same result for identifying visually-relevant claims ($< 4-5\%$). We also experiment with the combination of *BERT* features with *CLIP*'s image features, which improves the overall accuracy further but indicates that the model relies strongly on text (65.8 vs. 57.7 visual retrieval %) rather than the combination. The stronger reliance on text is possibly not a trait of the model alone, but could be also caused by an incompatibility of *BERT* and CLIP$_I$ features.

Finally, we achieve the best performance (by $1-4\%$) on binary and tertiary (when trained on T) claim detection by fine-tuning the *ALBEF* with and without OCR, respectively (Table 3, block 3,

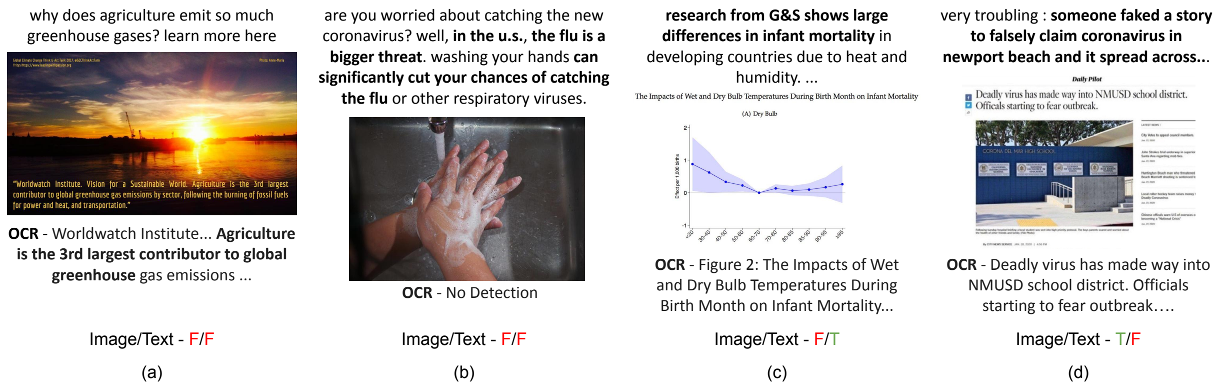


Figure 2: Qualitative examples where our best multimodal model classifies correctly and unimodal models do not. *F* - false classification, *T* - true classification.

last row). While the benefit of using OCR text in SVM models is not optimal and not considerably helpful, OCR addition to *ALBEF* retrieves the maximum number of visually-relevant claims (71.2%) without losing much on visually-irrelevant claims (79.3%) when trained on T (Table 4, block 2, last row). These results point towards a major challenge of combining multiple modalities and retaining intra-modal information (and influence) for the task at hand. As noted in section 4.3.1, an interesting result is that *ALBEF* in particular is less robust to resolved conflicts (split T_C) in the data when compared to just using *BERT*. On closer inspection, these conflicts are mostly caused by the image relevance to the text. The gap is further exaggerated in Table 6, where *ALBEF* performs much better than *BERT*, when conflict examples are removed from both training and evaluation. Figure 2 shows a few examples where our best multimodal model correctly classifies, whereas unimodal models based on either image or text do not. All the samples in the figure have images that have some connection to the tweet text. The image in Figure 2b has a connection to one of the words or phrases (e.g., washing your hands) in the tweet text but is not relevant for the claim itself. Figure 2a includes an image with the claim itself and a very generic scene in the background. Both image and text in Figure 2c and Figure 2d are relevant, and the image acts as evidence and additional information. In all these examples, a rich set of information extraction and complex cross-modal learning is required to identify claims in multimodal tweets. When comparing results of recent state-of-the-art architectures for fake news detection, SpotFake (Singhal et al., 2019) does considerably better than MVAE (Khattar et al., 2019) but worse than any of our baseline models.

5 Conclusions

In this paper, we have presented a novel *MM-Claims* dataset to foster research on multimodal claim analysis. The dataset has been curated from Twitter data and contains more than 3000 manually annotated tweets for three tasks related to claim detection across three topics, *COVID-19*, *Climate Change*, and *Technology*. We have evaluated several baseline approaches and compared them against two state-of-the-art fake news detection approaches. Our experimental results suggest that the fine-tuning of pre-trained multimodal and unimodal architectures such as *ALBEF* and *BERT* yield the best performance. We also observed that the overlaid text in images is important in information dissemination, particularly for claim detection. To this end, we evaluated a couple of strategies to incorporate OCR text into our models, which yielded a much better trade-off between identifying visually-relevant and visually-irrelevant (text-only) claims.

In the future, we will explore other and novel architectures for multimodal representation learning and other information extraction techniques to incorporate individual modalities better. We also plan to investigate fine-grained overlaps of concepts and meaning in image and text, and expand the dataset to COVID-19 related sub-topics and specific climate change events.

Acknowledgements

This work was funded by European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no 812997 (CLEOPATRA project), and by the German Federal Ministry of Education and Research (BMBF, FakeNarratives project, no. 16KIS1517).

References

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2018. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th international conference on social media and society*, pages 226–230.

Firoj Alam, Shaden Shaar, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2020. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *CoRR*, abs/2005.00033.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale COVID-19 twitter chatter dataset for open scientific research - an international collaboration. *CoRR*, abs/2004.03688.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. 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*, pages 215–236. Springer.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 747–754. The Association for Computer Linguistics.

Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Stuart E. Middleton, Andreas Petlund, and Yiannis Kompatsiaris. 2016. Verifying multimedia use at mediaeval 2016. In *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016*, volume 1739 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. 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)*, pages 558–563. Association for Computational Linguistics.

Gullal S. Cheema, Sherzod Hakimov, and Ralph Ewerth. 2020a. 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*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Gullal S Cheema, Sherzod Hakimov, and Ralph Ewerth. 2020b. Tib's visual analytics group at mediaeval'20: Detecting fake news on corona virus and 5g conspiracy. *MediaEval 2020 Workshop*.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021. On the role of images for analyzing claims in social media. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), Ljubljana, Slovenia, April 12, 2021 (online event due to COVID-19 outbreak)*, volume 2829 of *CEUR Workshop Proceedings*, pages 32–46. CEUR-WS.org.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *Scientific Reports*, 10(1):1–10.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

DGC. 2020. Un tackles 'infodemic' of misinformation and cybercrime in covid-19 crisis.

Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. 2020. Tweetscov19 - A knowledge base of semantically annotated tweets about the COVID-19 pandemic. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2991–2998. ACM.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. 2018. Tweetskb: A public and large-scale RDF corpus of annotated tweets. In *The Semantic Web - 15th International Conference, ESWC 2018,* *Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Fayez. 2021. A simple, pillow-friendly, wrapper around the tesseract-ocr api for optical character recognition (ocr).

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654. IEEE.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Lesa: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. *arXiv preprint arXiv:2101.11891*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA*, pages 770–778. IEEE Computer Society.

Neslihan Iskender, Robin Schaefer, Tim Polzehl, and Sebastian Möller. 2021. Argument mining in tweets: Comparing crowd and expert annotations for automated claim and evidence detection. In *Natural Language Processing and Information Systems - 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23-25, 2021, Proceedings*, volume 12801 of *Lecture Notes in Computer Science*, pages 275–288. Springer.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa, and Tanmoy Chakraborty. 2020. Newsbag: A benchmark multimodal dataset for fake news detection. In *Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020, New York City, NY, USA, February 7, 2020*, volume 2560 of *CEUR Workshop Proceedings*, pages 138–145. CEUR-WS.org.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA*, pages 2915–2921. ACM.

Hannah Rose Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki Markus Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. *CoRR*, abs/2107.04313.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Rabindra Lamsal. 2020. Coronavirus (covid-19) tweets dataset.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500. Dublin City University and Association for Computational Linguistics.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina*, pages 185–191. AAAI Press.

Justin Littman and Laura Wrubel. 2019. Climate Change Tweets Ids.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157. European Language Resources Association.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The CLEF-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 639–649. Springer.

Letitia Parcalabescu, Nils Trost, and Anette Frank. 2021. What is multimodality? In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*.

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sara Rosenthal and Kathleen R. McKeown. 2012. Detecting opinionated claims in online discussions. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012*, pages 30–37. IEEE Computer Society.

Alessandro Rovetta and Akshaya Srikanth Bhagavathula. 2020. Covid-19-related web search behaviors and infodemic attitudes in italy: Infodemiological study. *JMIR public health and surveillance*, 6(2):e19374.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Celia Andreu Sánchez and Miguel Angel Martín Pascual. 2020. Fake images of the sars-cov-2 coronavirus in the communication of information at the beginning of the first covid-19 pandemic. *El profesional de la información*, 29(3):6.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *Fifth IEEE International Conference on Multimedia Big Data, BigMM 2019, Singapore, September 11-13, 2019*, pages 39–47. IEEE.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK*, pages 849–857. ACM.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 540–547.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021*, pages 131–142. Association for Computational Linguistics.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. 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)*, pages 2099–2108. Association for Computational Linguistics.

A Appendix

In the following we include additional hyper-parameter (A.1) details and experimental results (A.2), additional dataset and annotation process details (A.3), and some annotated tweets for multimodal claim detection (A.4).

A.1 Other Hyper-paramter Details

For fine-tuning *BERT* and *ALBEF*, we use a batch size of 16 and 8 (size constraints), respectively. We train the models for five epochs and use the best performing model (in terms of accuracy on the validation set) for evaluation. For *BERT*, a dropout with the ratio of 0.2 is applied before the classification head. Further, we use AdamW (Loshchilov and Hutter, 2019) as the optimizer with a learning rate of $3e-5$ and a linear warmup schedule. The learning rate is first linearly increased from 0 to $3e-5$ for iterations in the first epoch and then linearly decreased to 0 for the rest of the iterations in 4 epochs. For *ALBEF*, we use the recommended fine-tuning hyper-parameters and settings from the publicly available code.

A.2 Additional Experimental Results

A.2.1 CLIP Variants

We experiment with *CLIP*'s three variants that use different visual encoder backbones, ResNet-50 (RN50), ResNet-50x4 (RN504) and a vision transformer (ViT-B/16) (Dosovitskiy et al., 2021) with *BERT* as textual encoder backbone. We select the models for textual and multimodal SVM experiments based on the performance (higher accuracy) using features from the visual encoders. Table 5 shows different visual encoders' features (with SVM) performance on binary and tertiary claim detection.

It should be noted that just like *ALBEF*, *CLIP* models can be fine-tuned with image-text tweet pairs for binary and tertiary tasks. However, when we experimented with fine-tuning the last few layers of *CLIP* with a classification head on top, it always performed worse than using extracted features for classification with SVM. This phenomenon is probably because of our relatively smaller sized labeled dataset, which is not enough for fine-tuning *CLIP* for the task.

Task $\rightarrow$	**BCD**		**TCD**			
Data Splits $\rightarrow$	$T_C \rightarrow E_C$		$T \rightarrow E_C$		$T_C \rightarrow E_C$	
Models $\downarrow$	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
RN50	66.3	65.7	64.1	50.6	**62.4**	**48.7**
RN50x4	**70.0**	**69.9**	61.5	**51.5**	61.4	48.5
ViT-B/16	68.6	68.4	**64.3**	49.8	59.7	48.3

Table 5: CLIP's different visual encoder backbones features' performance evaluation. Accuracy (Acc) and Macro-F1 (F1) for binary (BCD) and tertiary claim detection (TCD) in percent [%]. As described in Section 3.5, we use the Training Split (T) and Evaluation (Testing) Split (E) with resolved (index C) and without (no index) conflicts.

A.2.2 Results for "without conflicts" (E) Evaluation Split

In Section 4, we show results for tertiary claim detection (TCD) on evaluation splits "with resolved conflicts" (E_C) by training on T and T_C. Here in Table 6, we show the evaluation on "without conflicts" evaluation split (E). As with evaluation on E_C, multimodal models are more sensitive to training on T_C where conflict resolution strategy causes the accuracy to drop for all models. However, *CLIP* and *ALBEF* models, in this case, have higher F1-score (as well as accuracy) when trained on T. Even with less training data, the models perform better and best among all evaluated multimodal models. In the case of training on T_C, *BERT* performs the best, which is closely followed by *ALBEF* with OCR text.

As described in section 4.3.1, the evaluation of retrieved visually-relevant and visually-irrelevant claims on E follows the evaluation on E_C. Even though $CLIP_I$ and fine-tuned *BERT* retrieves the most amount of two types of claims, all models do better when trained on T_C than on T.

Overall, for a realistic scenario, training on T_C gives the best performance trade-off between Acc, F1 and retrieved claims for multimodal models.

A.2.3 Confusion Matrix

Following the results on E_C in section 4 for binary and tertiary tasks, we show normalized (by row) confusion matrices based on predictions from the $ALBEF \oplus OCR \oplus FT$ model. Figure 3a is the confusion matrix on E_C for binary claim detection (BCD). Whereas, Figure 3b shows the matrices on E_C with training on T_C (b.1) and T (b.2). Although the not-claim's true positives remain the same, confusion for the not-check-worthy and check-worthy class is less severe when trained on T_C.

Task →	TCD			
Data Splits →	$T \rightarrow E$		$T_C \rightarrow E$	
Models ↓	**Acc**	**F1**	**Acc**	**F1**
Random	33.7	28.2	33.7	28.2
Majority	62.7	38.5	62.7	38.5
ImageNet	62.5	40.9	62.5	42.1
CLIP$_I$	68.9	50.2	67.2	48.7
BERT	77.9	52.9	72.8	56.9
↪ FT	78.3	51.2	**79.2**	**61.4**
CLIP$_T$	77.3	54.4	71.6	52.3
BERT ⊕ ImageNet	77.5	56.0	77.0	56.9
↪ ⊕ OCR	77.7	55.0	76.6	55.8
CLIP$_{I \oplus T}$	77.5	56.4	73.0	52.6
CLIP$_I$ ⊕ BERT	77.9	53.3	72.6	56.8
ALBEF	76.6	55.0	67.6	52.7
↪ FT	**80.0**	63.3	76.8	59.7
↪ ⊕ OCR ⊕ FT	78.7	**63.5**	77.5	59.9
MVAE	64.8	40.7	62.9	43.2
SpotFake	72.8	49.7	70.7	50.4

Table 6: Accuracy (Acc) and Macro-F1 (F1) for tertiary claim detection (TCD) in percent [%]. As described in Section 3.5, we use the Training Split (T) and Evaluation (Testing) Split (E) with resolved (index C) and without (no index) conflicts. Additional results on evaluation split without conflicts (E). Unless FT (fine-tuning) is written, all models (except MVAE and SpotFake) are SVM models trained on extracted features.

Data Splits →	$T \rightarrow E$		$T_C \rightarrow E$	
Models ↓	**V (76)**	**T (120)**	**V (76)**	**T (120)**
ImageNet	39.8	39.2	67.1	58.3
CLIP$_I$	72.4	69.2	**78.9**	76.7
BERT	52.6	80.0	61.8	85.0
↪ FT	53.9	79.2	57.9	**85.8**
CLIP$_T$	51.3	76.7	60.5	76.7
BERT ⊕ ImageNet	63.2	68.3	75.0	80.8
↪ ⊕ OCR	69.7	78.3	75.0	81.7
CLIP$_{I \oplus T}$	68.4	70.0	76.3	78.3
CLIP$_I$ ⊕ $BERT$	60.5	75.0	60.5	85.0
ALBEF	63.2	77.5	65.8	80.8
↪ FT	65.8	79.2	75.0	80.8
↪ ⊕ OCR ⊕ FT	**76.3**	**82.5**	77.6	85.0

Table 7: Visually-relevant (V) and visually-irrelevant (T) claim detection evaluation. The amount of test samples is reported in brackets and the fraction, how many of them were retrieved, is given in percent [%]. Additional results on evaluation split without conflicts (E). The underlying models are trained for binary claim detection (BCD). The labels for visual relevance are only used for retrieval evaluation.

A.2.4 Ablation on OCR length

The amount of text that can be detected from an image varies, as it can be seen in Figure 8. As a consequence, we experimented with the length of OCR text in terms of the number of words for both binary and tertiary claim detection with *ALBEF*. We observe (see Figure 5) that 128 words give

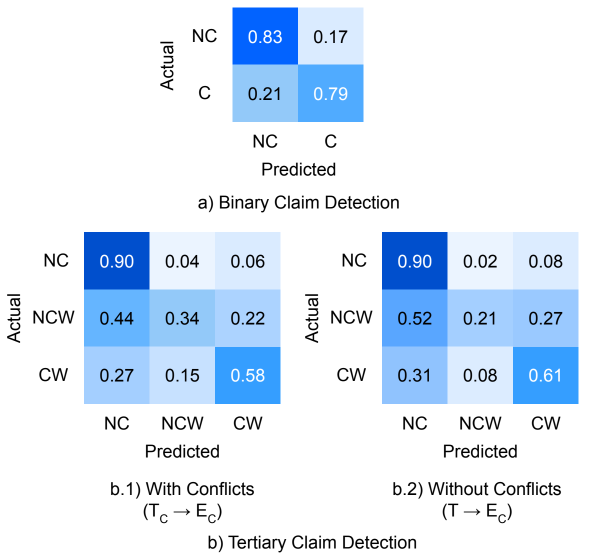


Figure 3: Normalized (by row) Confusion Matrices for the Binary and Tertiary Claim Classification Tasks. NC: Not-Claim, NCW: Not-check-worthy-Claim, C: Claim, CW: check-worthy-Claim

comparable or better performance than any less number of words in OCR text across tasks and number of layers fine-tuned. We chose 128 words instead of 64 because the model with 128 words showed a balanced performance for binary, tertiary and retrieved claims. Models with 64 or greater than 128 words had a lower performance for either visually-relevant or irrelevant retrieved claims.

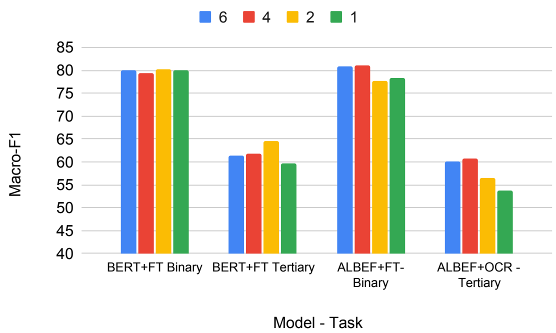


Figure 4: Ablation experiment on number of layers fine-tuned in *BERT* and *ALBEF*

A.2.5 Ablation on number of layers trained

We ran ablation experiments to see the effect of training the last few layers of *BERT* and *ALBEF* ⊕ OCR. We experiment with fine-tuning the last six, four, two layers and only the last layer of each model. The results are shown in Figure 4. Overall, fine-tuning the last two and four layers of *BERT* and *ALBEF* respectively gives the best results. Therefore, all the fine-tuning results for *BERT*, *ALBEF* and *ALBEF* $\oplus$ OCR are based on the above observation. For fine-tuning six or more layers, the unlabeled dataset can be incorporated in the future as a pre-training step followed by task-specific training.

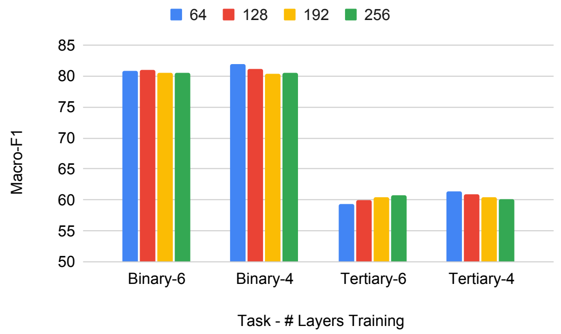


Figure 5: Ablation experiment on OCR text length (number of words) in *ALBEF*

A.3 Additional Dataset and Annotation Details

A.3.1 Claim Definition

Factually-verifiable Claims: should ideally have some of the following information (extended from Barrón-Cedeno et al. (2020)):

- reference to who, where, when, what, etc
- a definition, procedure, law or a process
- numbers or quantities in the tweet, e.g. sums of money, number of cases or deaths
- verifiable predictions
- refers to people, events, (event) locations
- refers to images and videos in the tweet
- personal opinions with claims that have factually-verifiable information

Check-worthy Claims: We follow a similar definition as Barrón-Cedeno et al. (2020), where claims are check-worthy if the information has some of the following properties:

- *Harmful*: if the statement attacks a person, organization, country, group, race, community, etc. The intention of such statements can be to spread rumours about an individual or a group, which should be checked by a professional or flagged and prioritized for further checking.

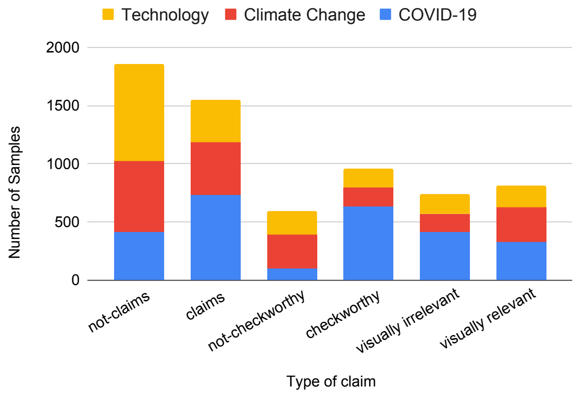


Figure 6: Class distributions in the annotated dataset ("with resolved conflicts") across different topics

- *Urgent or breaking news*: such statements are news-like where the claim is about prominent people (public personality like politicians, celebrities), organizations, countries and events (like disease outbreaks, forest fires, stock market crash).
- *Up-to-date*: such claims often refer to official documents and contain parts of clauses in climate agreements or articles in a constitution. This information is vital for checking, as many people consume social media as means of news, information and believe it to be true.

A.3.2 Filtering Strategies

The following Table 8 shows number of samples after each filtering step. The duplicate removal is performed across all the data irrespective of the topic in order to avoid duplicates that might fall into more than one topic.

Filtering Strategy	COVID	Climate	Tech.
No Filter	214 715	28 374	417 403
Empty text	214 715	28 374	417 403
Duplicate removal	28 522	11 333	383 043
Tweets with no image	28 522	11 333	383 043
Text not in English	28 148	11 274	377 532
Image size (200x200)	27 572	10 895	369 735
Hashtags > 6	26 786	10 013	287 242
Top-300 Hashtags	17 771	4874	62 887

Table 8: Data corpus statistics after applying different filtering strategies (in order).

A.3.3 Class Distributions Across Topics

In Figure 6, we provided the topic and class distributions in the labeled dataset.

A.3.4 Split-wise Statistics

The following Table 9 shows split-wise distribution of topics and labels in data. Numbers in red and black are for "with resolved conflicts" and "without conflicts" splits, respectively.

Types of Labels	COVID	Climate	Tech
Not Claims	306/34/73 306/34/73	449/38/120 449/38/120	617/81/136 617/81/136
Claims	545/64/123 478/58/104	351/35/70 251/24/48	265/30/63 198/21/44
Not check-worthy	77/8/16 25/4/3	238/27/23 141/16/5	155/24/24 97/9/8
check-worthy	468/56/107 453/54/101	113/8/47 110/8/43	110/6/39 101/12/36
Not Visual	302/31/78 285/30/70	112/8/33 91/10/21	125/15/34 104/10/29
Visual	243/33/45 193/28/34	239/27/37 160/14/27	140/15/29 94/11/15
Total	851/98/196 784/92/177	800/73/190 700/62/168	882/111/199 815/102/180

Table 9: Labeled data characteristics in terms of type of labels and topic. Shown as Training/Validation/Evaluation splits. Second and third blocks are claims which are check-worthy (and not) and visual claims (and not) respectively. Red - "with resolved conflicts" and black - "without conflicts"

A.3.5 Relevant Hashtags

Although we crawl tweets from topic-based corpora, we further filter tweets by manually marking top 300 hashtags (sorted by occurrence) relevant to the topic. Figure 7 shows top-20 relevant hashtags for each topic.

A.3.6 Annotation Tool

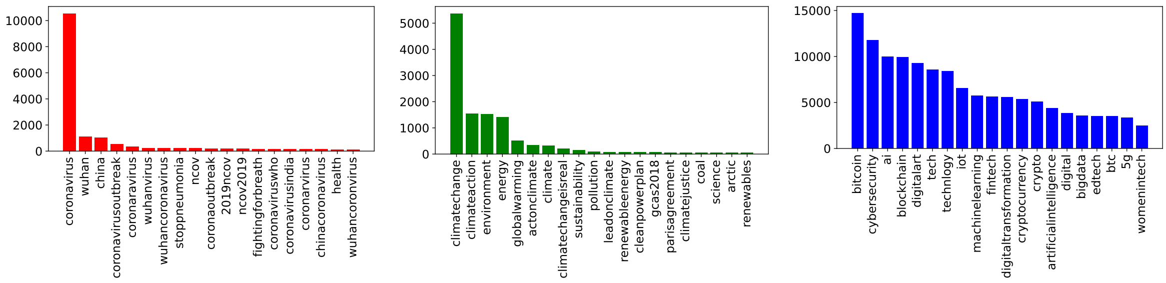
Figure 7d shows the annotation screen with the image-text pair, claim questions and a text box for feedback on difficult and missing image tweets.

A.4 Annotated Samples from the MM-Claims Dataset

We included multiple annotated samples corresponding to *visually-relevant claim* (see Figure 8) and *not a claim* (see Figure 9) classes.

Figure 7: Top-20 manually selected hashtags for topic relevance filtering strategy.

(a) COVID-19 (b) Climate Change (c) Technology



(d) Graphical User Interface that is used to annotate image-text tweets

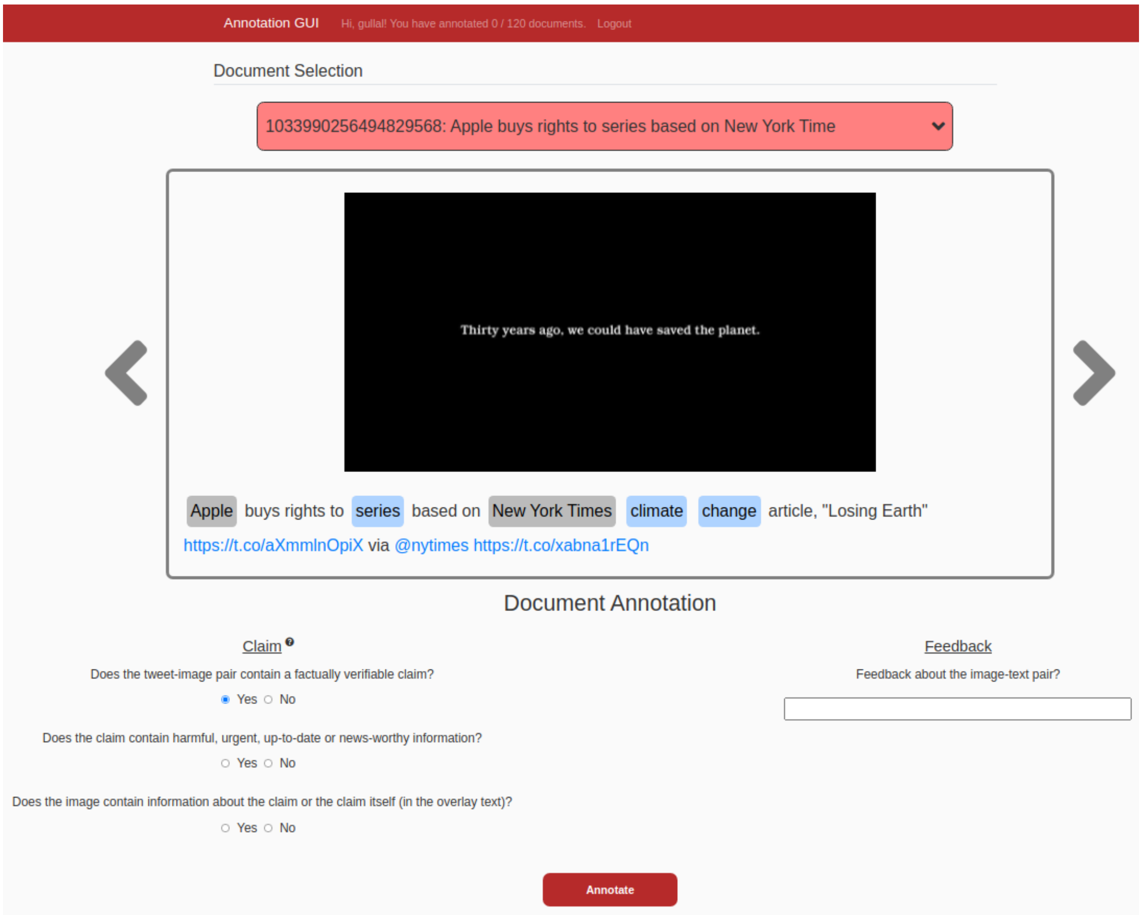


Figure 8: Additional examples for visually relevant claims for the topics *COVID-19* (bottom row), *Climate Change* (middle row), and *Technology* (top row).

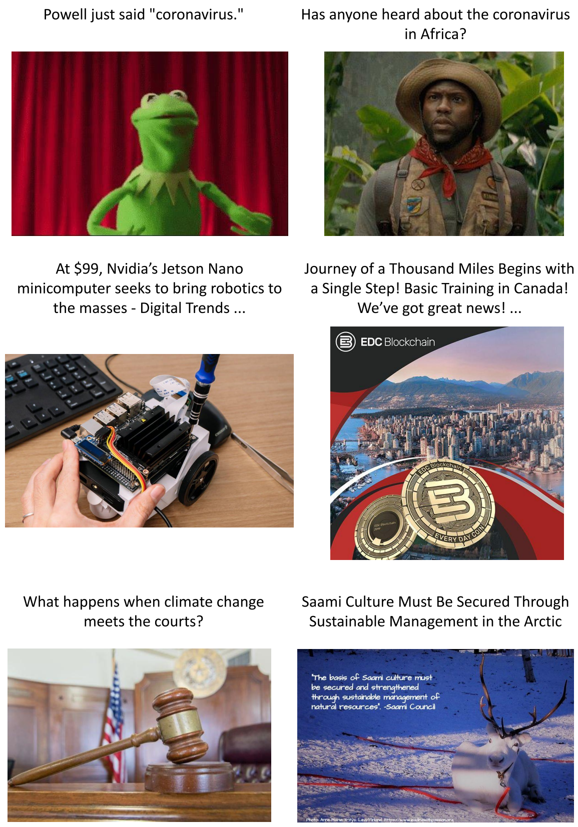


Figure 9: Additional examples that are not-claims for the topics *COVID-19* (top row), *Climate Change* (bottom row), and *Technology* (middle row).