FACTIFY: A Multi-Modal Fact Verification Dataset

Shreyash Mishra[1], S Suryavardan[1], Amrit Bhaskar[2], Parul Chopra[3], Aishwarya Reganti[4], Parth Patwa[5], Amitava Das[6,7], Tanmoy Chakraborty[8], Amit Sheth[7], Asif Ekbal[9] and Chaitanya Ahuja[3]

[1]*IIIT Sri City, India*
[2]*Arizona State University, USA*
[3]*Carnegie Mellon University, USA*
[4]*Amazon, USA*
[5]*University of California Los Angeles, USA*
[6]*Wipro AI labs, India*
[7]*AI Institute, University of South Carolina, USA*
[8]*IIIT Delhi, India*
[9]*IIT Patna, India*

Abstract

Combating fake news is one of the burning societal crisis. It is difficult to expose false claims before they create a lot of damage. Automatic fact/claim verification has recently become a topic of interest among diverse research communities. Forums like FEVER, FNC [1, 2] aim to discuss automatic fact-checking on text. Research efforts and datasets on text fact verification could be found, but there is not much attention towards multi-modal or cross-modal fact-verification. In order to bring the attention of the research community towards understanding multimodal misinformation, we release a multimodal fact checking dataset named FACTIFY. It is notably the largest multimodal fact verification public dataset consisting of 50K data points, covering news from India and the US. FACTIFY contains images, textual claims, reference textual documents and images labeled with three broad categories namely - *support, no-evidence, and refute.*

Keywords

Fake News, Fact Verification, Multimodality, Dataset, Machine Learning, Entailment

De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada

shreyash.m19@iiits.in (S. Mishra); suryavardan.s19@iiits.in (S. Suryavardan); abhask10@asu.edu (A. Bhaskar); parulcho@andrew.cmu.edu (P. Chopra); amitava.das2@wipro.com (A. Das)

1. Introduction

In the recent years, automatic fact checking has emerged to be an important problem in the AI community, since dangers of fraudulent claims masquerading as declarations of reality have become common. Although the birth of this problem goes back to the initial years of printing press, it has attracted increasing interest with the usage of social media. The rapid distribution of news across numerous media sources has resulted in the fast development of erroneous and fake content. It is tough to uncover misleading statements before they cause significant harm. According to statistics [3], about 67% of the American population believes that fake news produces a lot of uncertainty, and 10% of them knowingly propagate fake news. On the contrary, only 26% of respondents said they feel confidence in their ability to recognize bogus news.

The scarcity of available training data has been a fundamental obstacle in automated fact-checking. Recently, significant progress has been made with the release of two of the largest datasets - FEVER [1] and LIAR [4], among several others. LIAR contains 12.8K claims along with their meta-data (i.e., speaker of the claim, political affiliations of the speaker, medium through which the claim was first published) collected from the real fact-checking websites like Politifcat. Huge advancements have been achieved since the release of LIAR. A significantly larger dataset - FEVER includes proof and extensive meta-data to contextualize the claims even more. FEVER consists of 185K claims which were manually curated from Wikipedia. Although FEVER is a large dataset, it was purpose-made for research and this limits its ability to capture patterns from the real-world data. We release a multimodal fact checking dataset, called FACTIFY, which would aid in resolving this problem as it consists of original samples with no post-processing or manual data creation involved. Additionally, the visual cues that support textual claims would help the system to detect fake content with greater confidence. The dataset is released at https://competitions.codalab.org/competitions/35153 and the baselines are available at https://github.com/Shreyashm16/Factify .

Although there are research initiatives [5, 6, 7] and datasets [1, 4], on textual fact verification, there is less focus on multi-modal or cross-modal fact verification. The majority of the present fact-checking research relies on unimodal techniques, synthetic data production, and limited annotated datasets. Therefore, we believe that FACTIFY can serve as a stepping stone to build novel multimodal fact verification systems. The dataset contains images, textual claim, reference textual document/image. The task is to tag support, no-evidence, and refute between given claims; each of these categories are explained in the next section. The first two categories are further sub-divided into text and multimodal components. Thus, in total, all the data samples are labeled with one out of five choices. We choose twitter handles of popular news channels from the two large nations – the US and India. Therefore, the dataset entirely consists of real samples gathered from different social media news handles popular in India and the US.

To summarize, in this paper, we release a novel multimodal fact-checking dataset that can be used as a benchmark for researchers. We also propose unimodal and multimodal baseline models for our dataset. The paper is organised as follows: The proposed task is described in Section 2. Related work is described in Section 3. Data collection and data distribution are explained in Section 4 while Section 5 demonstrates the baseline model. Section 6 shows the results of our baseline models. Finally, we summarise our task along with the further scope and open-ended pointers in Section 7.

2. The Factify Task

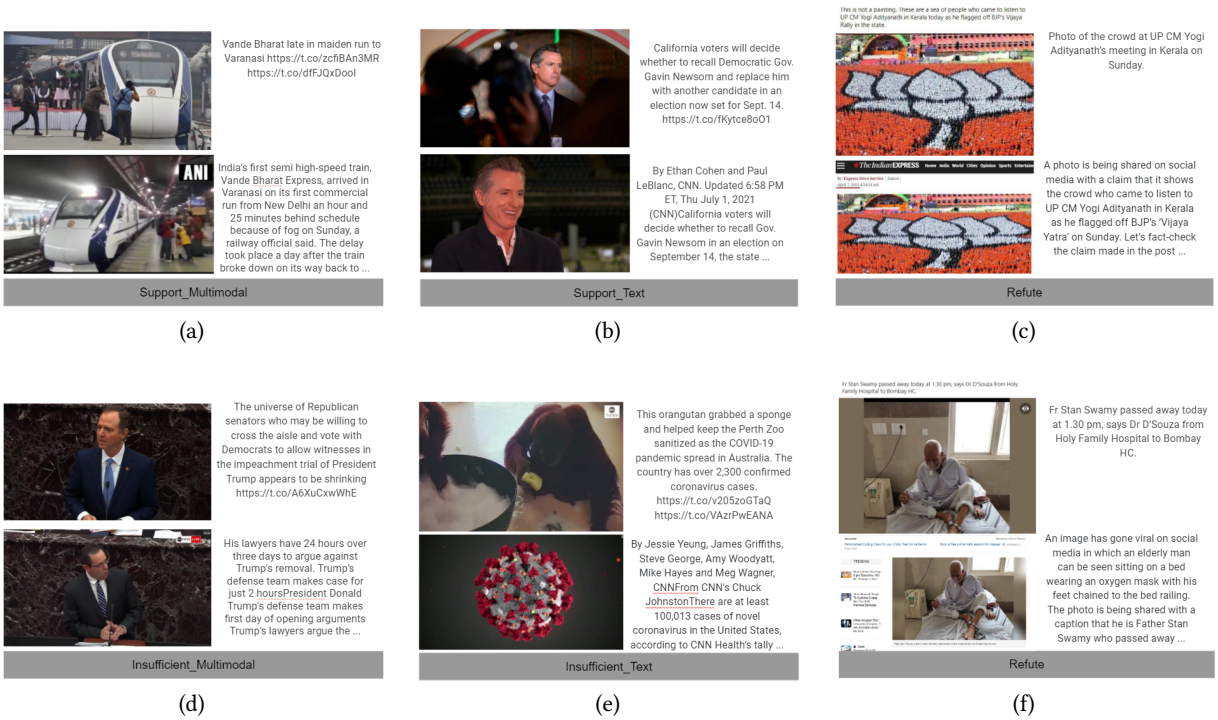

Figure 1: These are examples for all the 5 categories. The 'Multimodal' category i.e. (a) and (d) have similar images whereas 'Text' i.e. (b) and (e) have dissimilar images. The 'Support' category indicates that the given document supports the claim, as shown by examples (a) and (b). The 'Insufficient' category indicates that there is not enough information in the document to support or reject the claim, as shown by examples (d) and (e). Finally, the examples (c) and (f) are from the 'Refute' category. These are the false claims along with their supporting documents.

To detect multimodal fake news, we model the task as a multimodal entailment. We assume that each data point contains a reliable source of information, called "document", and its associated image and another source whose validity must be assessed, called the "claim" which also contains a respective image. The goal is to identify if the claim entails the document. Since we are interested in a multimodal scenario with both image and text, entailment has two verticals, namely textual entailment and visual entailment and their respective combinations. This data format is a stepping stone for the fact checking problem where we have one reliable source of news and want to identify the fake/real claims given a large set of multimodal claims. Therefore the task essentially is – given a textual claim, claim image, document and document image, the system has to classify the data sample into one of the five categories: *Support_Text*, *Support_Multimodal*, *Insufficient_Text*, *Insufficient_Multimodal* and *Refute*. The images are also supported by the text obtained by running an OCR. The descriptions of the labels are as follows-

- `Support_Text`: the claim text is similar or entailed but images of the document and claim

are not similar.

- `Support_Multimodal`: both the claim text and image are similar to that of the document.
- `Insufficient_Text`: both text and images of the claim are neither supported nor refuted by the document, although it is possible that the text claim has common words with the document text.
- `Insufficient_Multimodal`: the claim text is neither supported nor refuted by the document but images are similar to the document.
- `Refute`: The images and/or text from the claim and document are completely contradictory i.e, the claim is false/fake.

Figure 1 shows some examples of the classes.

Although we use this dataset for fake news detection, it is just one out of many applications of a bigger research problem - Multimodal entailment. Hence our dataset will serve a larger community.

3. Related Work

Over the last few years, various fact checking and fact verification datasets have been published. Majority of them being text based and only a few being multi-modal datasets. The textual datasets can broadly be grouped into two categories based on the information they provide.

The first category includes datasets that aim to predict the veracity based on the claim alone. LIAR [4] contains 12.8k manually labeled claims from politifact with 6 fine-grained labels and metadata such as speaker name. CREDBANK [8] focuses on checking credibility by providing tweets related to 1k events with manual credibility annotation. The Lie Detector dataset [9] approaches the task with 'true' and 'deceptive' text samples of size 600. Another such dataset uses Claim Matching [10] and has 2k pairs of multi-lingual text with labels based on text pair similarity. A dataset on Covid-19 fake news is provided by [11].

The second category includes datasets where the claim is accompanied with documents annotated with labels indicating whether the document supports the claim or is unrelated to it. A very well known dataset of this type is FEVER [1]. It contains 185k samples with a claim and a supporting document from Wikipedia, but, these claims were manually generated and then altered before being classified as '*Support*', '*Refute*' or '*NotEnoughInfo*'. MultiFC [12] is a multi-domain dataset of size 35k with claims and rich-metadata from 26 different websites. It has a wide range of labels preserved from these websites such as '*correct*', '*incorrect*', '*mis-attributed*' and '*not the whole story*'.

Textual datasets are no longer enough in the social media age. It is important to consider both the image and text when detecting fake news. Fakeddit [13] is a multi-modal dataset providing an image associated with a text. The image can be used as evidence for the text or vice-versa. Each of its 1 million samples has both high-level and fine-grained labels. It is similar to a image-caption dataset, which could result in a disjoint claim and image. FakeNewsNet [14] contains 23k articles with context and spatio-temporal information focused on fake news source and mitigation. The data and their labels have been obtained from fact checking websites such as Politifact and GossipCop. A dataset of fact-checked images shared on WhatsApp

during the 2018 Brazilian and 2019 Indian Elections [15] provides two sets of 135 and 897 images containing misinformation from Brazil and India, respectively. These fact-checked fake images from WhatsApp are supported by data from fact checking websites and manual expert annotations. Table 1 summarizes datasets and their statistics. To the best of our knowledge, Factify is the largest real-world multimodal fake news detection dataset. The dataset has five categories based on the entailment of the text and image pairs. It supports the automation of fact checking using an entailment approach.

Name	# Claims	# Labels	Data	Year
LIAR [4]	12836	6	Claim Text, Metadata (Speaker etc.)	2017
CREDBANK [8]	1049	5	Claim Text, Event, Topic	2015
The Lie Detector [9]	600	2	Claim Text	2009
Claim matching beyond english [10]	2343	3	Claim Text Pairs	2021
FEVER [1]	185445	3	Claim Text, Document Text	2018
MultiFC [12]	36534	40	Claim Text, Document url, Metadata	2019
Fakeddit [13]	1 million	2/3/6	Claim Text, Claim image	2019
Covid-19 Fake News dataset [11]	10700	2	Claim Text	2020
FakeNewsNet [14]	23921	2	Claim Text, Spatiotemporal info	2019
Whatsapp fact-checking dataset [15]	1032	3	Claim Image, Metadata	2020
Factify (ours)	50000	5	Claim Text, Claim Image, Document Text, Document Image, Images OCR	2021

Table 1
Details of related public datasets for automated fact-checking along with available meta data and release year.

4. Data

4.1. Data Collection

We collected date-wise tweets from twitter handles of Indian and US news sources: Hindustan Times [1], ANI[2] for India and ABC[3], CNN [4] for US based on accessibility, popularity and posts per day. Moreover, these twitter handles are eminent for their objective and disinterested approach. From each tweet, we extracted the tweet text and the tweet image(s). Now, for each tweet, we do the following:

[1]https://twitter.com/htTweets
[2]https://twitter.com/ANI
[3]https://twitter.com/ABC
[4]https://twitter.com/CNN

- For each tweet of account A, we got similar tweets from account B. Similarity is measured on the basis of text. Text similarity is measured using Sentence BERT first, and then the extent of common words is measured as the second metric.
- Next, the the image similarity for the corresponding images of the tweet pair was calculated. Image similarity is measured using histogram similarity and cosine similarity on a pre-trained ResNet50 model.
- According to the scores for each of these measures, the tweet pair is classified into 4 categories: `Support_Multimodal`, `Support_Text`, `Insufficient_Multimodal` and `Insufficient_Text`. The various thresholds used for classification are listed in Figure 3.
- From this tweet pair, we selected a tweet (say tweet B) and obtained the url for the corresponding article published on the source's website from the tweet text. We then replaced the tweet text with article contents after scraping it (`document` in dataset). We do this so as to mimic real world fact checking process, i.e., manually comparing claims with documents or articles.
- The image OCRs were obtained using Google Cloud Vision API [5].

Here is the final description for each attribute in the dataset -

- `Claim`: Tweet A text
- `Claim_image`: Tweet A image
- `Claim_ocr`: Tweet A image OCR
- `Document`: Tweet B article text
- `Document_image`: Tweet B image
- `Document_ocr`: Tweet B image OCR
- `Category`

Figure 2 explains the five classes in our dataset.

For appropriate classification of the dataset, two similarity measures were computed.

Sentence Comparison: We use 2 methods to check similarity amongst sentences:

- Sentence BERT: Sentence BERT [16] is a modification of the BERT model that uses siamese and triplet network structures to get sentence embeddings. These sentence embeddings can be compared with each other to get their corresponding similarity score. We use cosine similarity as the textual similarity metric. We use Sentence BERT (SBERT) over the pre-trained BERT model and RoBERTa mainly because it is much faster without compromising the accuracy. For our application, we used the 'paraphrase-MiniLM-L6-v2[6]' pre-trained model. For each text pair, we derive their corresponding embeddings using the SBERT model, and check their cosine similarity. We manually decide on a threshold value *T1* for cosine similarity, and classify the text pair accordingly. If the cosine similarity score is greater than *T1*, then it is classified into the support category. On the other hand, if the cosine similarity score is lower than *T1*, the news may or may not be the same (the evidence at hand is insufficient to judge whether the news is same or not). Hence it is sent for another check before classifying it into Insufficient category.

[5] https://cloud.google.com/vision/docs/ocr

[6] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

Support_Multimodal	Text is supported Similar News	Image is supported
Support_Text	Text is supported Similar News	Image is neither supported nor refuted
Insufficient_Multimodal	Text is neither supported nor refuted May have common words	Image is supported
Insufficient_Text	Text is neither supported nor refuted May have common words	Image is neither supported nor refuted
Refute	Fake Claim	Fake Image

Figure 2: These are five categories in our dataset. Multimodal categories (Support Multimodal and Insufficient Multimodal) have similar images while Text only categories (Support Text and Insufficient Text) have similar words.

- NLTK: If the cosine similarity of the sentence pair is below T1, we use the NLTK library [17] to check for common words between the two sentences. If the common words score is above a different manually decided threshold *T2*, only then the news pair is classified into the insufficient category. Common words are being checked to ensure that the classification task is challenging. To check for common words, both texts in the pair are preprocessed, which included stemming and removing stopwords. The processed texts are then checked for common and similar words, and their corresponding scores are determined. If the common words score is greater than T2, the pair is classified as Insufficient else the pair is dropped.

Image Comparison: We use two metrics for determining whether images are similar or not:

- Histogram Similarity: The images are converted to normalized histogram format and similarity is measured using the correlation metric.
- Cosine Similarity: The images are converted to feature vectors using pre-trained ResNet50 model, and these feature vectors are used to calculate the cosine similarity score. Manually decided thresholds, as described in Figure 3, are used to judge whether the text and image pair is similar or not.

The text pairs are first classified into either Support or Insufficient categories, and then further sub-classified into *Support_Text*/ *Support_Multimodal* or *Insufficient_Text*/ *Insufficient_Multimodal* categories based on the similarity of the image pairs. If the corresponding images for the texts are similar, then they could be used to judge whether news is the same or not. The category where both the images and the texts are similar is called *Support_Multimodal*. The category where the images are similar but the texts were not is called *Insufficient_Multimodal*.

If the corresponding images for the texts were not similar, then they could not be used to judge whether news is the same or not. The category where both the images and the texts are not similar is called *Insufficient_Text*. The category where the texts are similar but the images are not is called *Support_Text*.

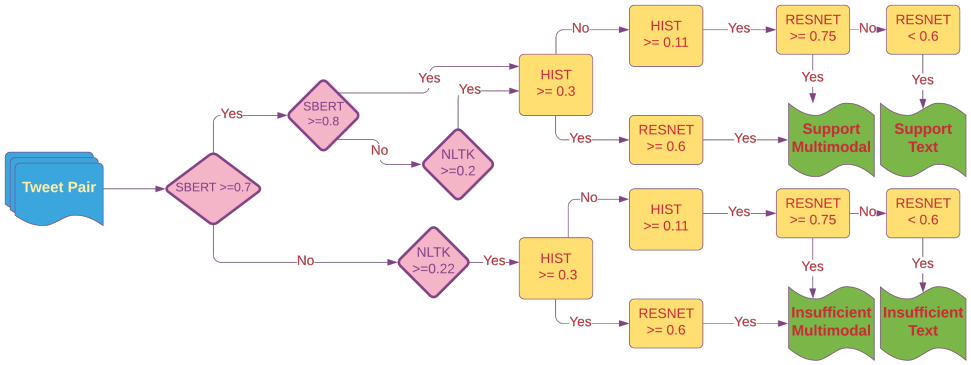

Figure 3: Text and image pair similarity based on classification thresholds on pre-trained models.

For the refute category, we scrape several reliable fact-check websites like Vishwas[7], Times of India[8], India Today[9], AFP India[10], AFP USA[11], AltNews[12], BOOM[13], Factly[14], NewsChecker[15], NewsMobile[16] and WebQoof[17]. For each article in these websites, we collect the claim (sentence that states the fake news), document (text that proves claim is false), claim images (fake news image, may be screenshot of fake-post), `document_image` (image is proof of fakeness of claim). The dataset is released at https://competitions.codalab.org/competitions/35153 .

4.2. Data Statistics And Analysis

In order to understand the nature and distribution, we provide preliminary analysis of the Factify dataset. The dataset has a total of 50000 samples, and each of the 5 categories has equal samples. The dataset has a Train-Val-Test split of 70:15:15 2.

To identify and predict the veracity of the claim, a common method is to collate a given claim and the corresponding news article or document. We analyze the word occurrence and distribution of the claims in Figure 4. We can observe that most fake news is related to politics and religion.

[7]https://www.vishvasnews.com
[8]https://timesofindia.indiatimes.com/times-fact-check
[9]https://www.indiatoday.in/fact-check
[10]https://factcheck.afp.com/afp-india
[11]https://factcheck.afp.com/afp-usa
[12]https://www.altnews.in/
[13]https://www.boomlive.in/fact-check
[14]https://factly.in/category/english/
[15]https://newschecker.in/
[16]https://newsmobile.in/
[17]https://www.thequint.com/news/webqoof

	Train	Validation	Test	Total
Support_Multimodal	7000	1500	1500	10000
Support_Text	7000	1500	1500	10000
Insufficient_Multimodal	7000	1500	1500	10000
Insufficient_Text	7000	1500	1500	10000
Refute	7000	1500	1500	10000
Total	35000	7500	7500	**50000**

Table 2
Dataset distribution statistics. Note that all the classes are balanced to eliminate data bias.

Entity	Frequency
minister	6978
president	6636
trump	5624
state	5322
cases	5111
new	5004
people	4474
#covid19	3939
video	3894
congress	3436
police	3410
first	3204
says	3142
chief	3076
india	3050
former	3004
bjp	2999
modi	2956
delhi	2942
indian	2842

Table 3
Top 20 most frequent words extracted from claim documents

The claims in the dataset are majorly associated with politics and governance. Claims from both the USA and India mention political parties and leaders, as shown by the top 20 most frequent entities listed in Table 3. The data captures other past or present affairs such as ”Covid19” aswell.

We show the number of unique n-grams for the Factify dataset in Table 4. This shows the lexical diversity of the dataset.

(a) Support (b) Insufficient (c) Refute

Figure 4: Word clouds indicating top words used in each class

N-gram	#	Examples
1-gram	61793	(case), (trump)
2-gram	307961	(prime,minister), (president,trump)
3-gram	411845	(prime,minister,narendra), (new,covid19,case)
4-gram	453139	(prime,minister,narendra,modi), (cast,vote,polling,booth)

Table 4
Unique n-grams for the claims in all categories

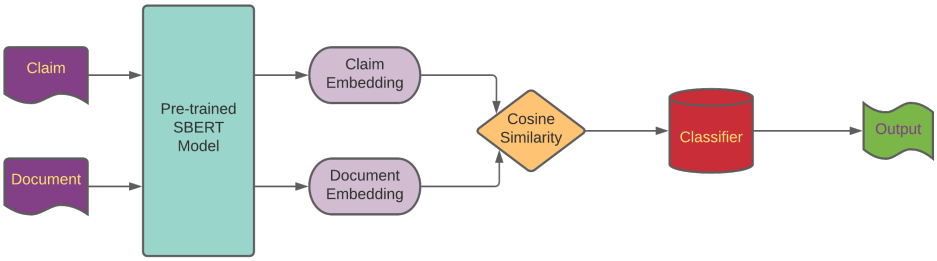

Figure 5: Text-Only Baseline Model which takes only claim text and document text as input

5. Baseline model

We explore 2 different settings to establish baselines i.e., text-only & multimodal. The goal is to identify the difference between using only one prime modality which is text and then augmenting image information to gauge the performance boost.

Text Only Model: This model (shown in figure 5) ignores the information given by the image. Instead of focusing on multimodal aspect of the data, this model focuses only on the textual aspect. To do so, the model creates sentence embeddings of claim and document attributes using a pretrained Sentence BERT model [16], 'paraphrase-MiniLM-L6-v2'. Then, cosine similarity

is measured on the embeddings. This score is used as the only feature for the dataset, and classification is performed using traditional machine learning classifiers like Support Vector Machine and Decision Tree.

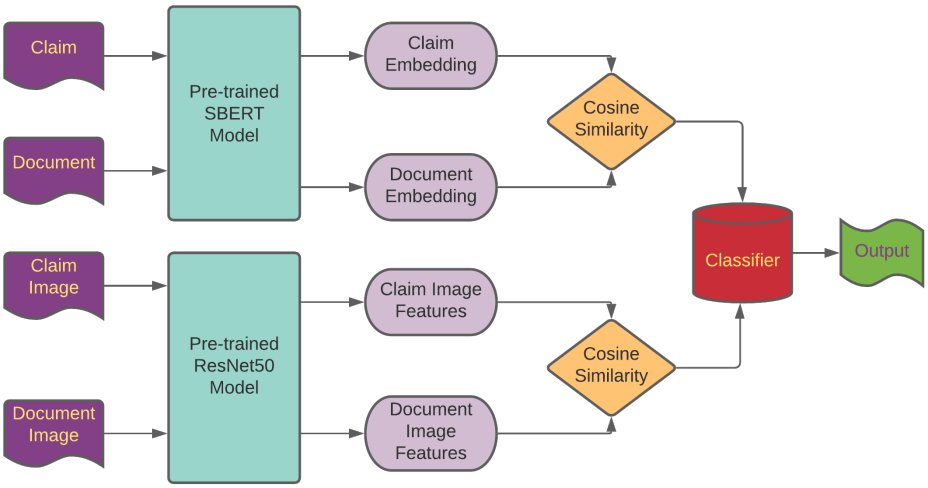

Figure 6: Multi-Modal Baseline Model which takes claim text, claim image as well as document text and document image as input.

Multi-Modal Model: Information shared online is very often of multi-modal nature. Images can change the context of a claim and lead to misinformation. Thus, it is important that we consider both the image and text when classifying the claims. As it is an entailment based approach, features from both claim and document image-text pairs must be extracted. This is done using the pre-trained ResNet50 model [18]. The cosine similarity score is computed between both the claim and document image features. The cosine similarity for the text embeddings is computed, same as the textual baseline model. The model diagram is shown in figure 6. The final classification F1 score is shown in the table 5 below for different classifiers trained on these two scores as attributes. There is an improvement in performance compared to the text-only model. The baselines are available at https://github.com/Shreyashm16/Factify .

6. Results

The results obtained for each of setting described above are presented in Table 5. We experiment with various classification models for both the text and multimodal settings. For the text only setting, our best performing decision tree model achieves an F1-Score of 41.3% on the test set. While in the multimodal setting achieves a best performance of 53.09%. Note that there is about ~9% performance improvement when image features are used, which suggests that the task performance heavily relies on multi-modal information. However, we use quite naive approaches to establish baselines to encourage more innovative approaches and there is a huge scope for improvement. The results also indicate that off-the-shelf models don't

perform very well on the task since the best performing model achieves only 53.09%. More comprehensive approaches like using vision-language pre-trained models, training on other related datasets/tasks and fine-tuning on Factify, innovative attention and fusion techniques will definitely boost performance. We leave such methods as future work.

Method	Algorithm	Validation Score	Test Score
Text-only	Logistic Regression	29.15%	29.14%
Text-only	KNN	37.20%	36.24%
Text-only	SVM	29.91%	29.88%
Text-only	Decision Tree	**42.53%**	**41.33%**
Text-only	Random Forest	36.18%	35.15%
Multimodal	Logistic Regression	49.96%	50.10%
Multimodal	KNN	47.17%	47.31%
Multimodal	SVM	52.32%	51.65%
Multimodal	Decision Tree	47.26%	49.37%
Multimodal	Random Forest	**54.11%**	**53.09%**

Table 5
Results of baseline machine learnig models. Note that multimodal models perform much better than unimodal models.

7. Conclusion and Future Work

In this work, we take a leap towards developing machine learning techniques for the multimodal fact verification by releasing a large real-world dataset with cues from two modalities namely text and image. We also release unimodal and multimodal baselines to emphasize on the difficulty of the problem and scope for improvement. However, our work only scratches the surface and many follow-up research directions can be pursued. In the current dataset, we assume that claims have a binary class i.e., either fake or true but there can be cases where the claim can be partially true or fake. We aim to incorporate these classes in our future work. We also envision to understand deeper relationships between text and image with the help of attention methods in the future.

References

[1] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).

[2] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, I. Gurevych, A retrospective analysis of the fake news challenge stance-detection task, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1859–1874. URL: https://aclanthology.org/C18-1158.

[3] A. Watson, Fake news in the u.s. - statistics & facts, Statista (2021).

[4] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).

[5] A. Hanselowski, H. Zhang, Z. Li, D. Sorokin, B. Schiller, C. Schulz, I. Gurevych, Ukp-athene: Multi-sentence textual entailment for claim verification, arXiv preprint arXiv:1809.01479 (2018).

[6] Z. Liu, C. Xiong, M. Sun, Z. Liu, Fine-grained fact verification with kernel graph attention network, arXiv preprint arXiv:1910.09796 (2019).

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

[8] T. Mitra, E. Gilbert, Credbank: A large-scale social media corpus with associated credibility annotations, in: ICWSM, 2015.

[9] R. Mihalcea, C. Strapparava, The lie detector: Explorations in the automatic recognition of deceptive language, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09, Association for Computational Linguistics, USA, 2009, p. 309–312.

[10] A. Kazemi, K. Garimella, D. Gaffney, S. A. Hale, Claim matching beyond english to scale global fact-checking, 2021. arXiv:2106.00853.

[11] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT) 2021, Springer, 2021, p. 21–29. URL: http://dx.doi.org/10.1007/978-3-030-73696-5_3. doi:10.1007/978-3-030-73696-5_3.

[12] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. Grue Simonsen, Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims, in: EMNLP, Association for Computational Linguistics, 2019.

[13] K. Nakamura, S. Levy, W. Y. Wang, r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, arXiv preprint arXiv:1911.03854 (2019).

[14] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media, 2019. arXiv:1809.01286.

[15] J. C. S. Reis, P. de Freitas Melo, K. Garimella, J. M. Almeida, D. Eckles, F. Benevenuto, A dataset of fact-checked images shared on whatsapp during the brazilian and indian

elections, 2020. arXiv:2005.02443.

[16] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[17] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.

[18] P. Kasnesis, R. Heartfield, X. Liang, et al., Transformer-based identification of stochastic information cascades in social networks using text and image similarity, in: Journal of Applied Soft Computing, 2021.