

Asatya at CheckThat! 2022: Multimodal BERT for Identifying Claims in Tweets

Manan Suri¹, Prajeet Katari¹ and Saumay Dudeja¹

¹Netaji Subhas University of Technology, New Delhi

Abstract

The paper presents an overview of our submission to the fifth edition of the CheckThat! Lab challenge. Specifically, our team participated in subtasks 1A, 1B and 1C under task 1, which aimed to identify relevant claims in tweets. More specifically, the three subtasks deal with evaluating the check-worthiness, presence of verifiable facts and presence of harmful content in the tweets, respectively. The lab provided us datasets for the three subtasks in multiple languages including English, Dutch and Bulgarian. A total of 14, 10 and 12 teams participated in the subtasks 1A, 1B and 1C respectively, out of which we ranked 9th, 2nd and 2nd, respectively.

In this paper, we discuss our methodology for the subtasks, which includes data augmentation to increase the size of our training dataset, followed by preprocessing of the tweets and feature extraction for the tweets from the Twitter API to gain more data points that can help gauge the credibility and/or authenticity of the tweet. Finally, we discuss the structure of our Multimodal model which uses numerical and categorical features in addition to the textual data from tweets.

Keywords

BERT, tweet classification, NLP, multimodal analysis, fake news, transformers, pre-trained models, Twitter API, data augmentation

1. Introduction

With the advent and uprise of social media, fake news, disinformation and the spread of propaganda have become some of the most pressing issues of the time. Although the spread of disinformation is not a new phenomenon, it has accelerated with time as people around the world have become more connected than ever, primarily through social media[1]. The spread of disinformation is an issue that affects almost all spheres of society and hence, several strategies for combating disinformation and its effects are currently being researched and implemented. However, they haven't proven to be sufficiently effective. Although no strategy can be perfect, the persisting prevalence of disinformation and fake news on social media indicates that there is still progress to be made in improving these strategies.

The CheckThat! Lab [2], part of the Conference and Labs of the Evaluation Forum (CLEF) was run for the fifth time in 2022 and aimed to foster technology that would help in identifying and dealing with disinformation present in tweets. Particularly, the tasks aim to evaluate the


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ manan.suri.ug20@nsut.ac.in (M. Suri); katari.ug20@nsut.ac.in (P. Katari); saumay.dudeja.ug20@nsut.ac.in (S. Dudeja)

🌐 <https://www.linkedin.com/in/manansuri27/> (M. Suri)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

check-worthiness, presence of verifiable claims, presence of content harmful to the society in the tweets, and further evaluate the amount of attention a tweet should get from the policymakers.

We participated in Task 1 of The CheckThat! Lab 2022 and completed 3 subtasks (1A, 1B, and 1C). We used a Multimodal model that uses BERT [3] for the textual data and further also includes categorical and numerical features for prediction. For the numerical and categorical features, we used certain data points relevant to a particular tweet (such as the credibility of the account, the links, and so on) extracted from the Twitter API to supplement the available knowledge. Finally, to expand our database and have an overall better trained model, we used data augmentation to increase the size of our dataset. This was performed by translating the Dutch and Bulgarian training sets. This translation was performed only in languages with similar scripts and semantics to ensure the preservation of information within the tweet.

Our team ranked #2 on subtasks 1B and 1C, and #9 on subtask 1A. In this paper, we shall describe our methodology in detail. The key advantages of our system were that it had a bigger dataset due to augmentation and the fact that we used not only textual but also categorical and numerical features to train our model.

Our code has been released and is publicly available.¹

2. Task Description

This year, the CheckThat![4] Lab campaign has a total of 3 tasks, each further divided into multiple subtasks. Our team participated in Task 1 [5], specifically in 3 subtasks in Task 1, namely 1A, 1B and 1C. Task 1 dealt with identifying relevant claims in tweets.

Subtask 1A: Subtask 1A aims to evaluate the check-worthiness of tweets. The goal of this task is to predict whether a given tweet is worth fact-checking. The task is offered in 6 languages, and our team provided a submission for the English test set.

Subtask 1B: Subtask 1B deals with detecting the presence of a verifiable claim in a given tweet. Given a tweet, this task requires us to predict whether it contains a verifiable factual claim. The task is offered in 5 languages, and our team submitted the English test set.

Subtask 1C: Subtask 1C deals with the evaluation of a tweet in terms of its effects on the consumers. Specifically, it aims to determine whether it harms the society. The task runs in 5 languages, and our team has provided a submission for the English test set during the evaluation phase.

All the three subtasks 1A, 1B, and 1C are binary classification tasks, with two labels: 0, 1 which represent a "No" and "Yes" response to the given specifics of the subtask. The datasets provided for all the three subtasks contain tweets in Arabic, Bulgarian, Dutch, English, and Turkish. Additionally, Spanish data is present for subtask 1A.

The metric used to evaluate the subtasks 1A and 1C is the F1 score, which is simply the harmonic mean of precision and recall and hence increases with improvement in both, precision and recall.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

¹<https://github.com/MananSuri27/MultimodalTweetAnalysis>

For subtask 1B, accuracy is used as the evaluation metric.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (2)$$

where TP, TN, FP, FN refer to True Positive, True Negative, False Positive, False Negative.

3. Data Description

The dataset for each of the languages in each of the subtasks was given in .tsv format with the following 5 columns:

- tweet_id – unique identifier of the tweets
- topic - topic to which the tweet is related.
- tweet_url – URL of the related tweet
- tweet_text – text content of the tweet
- class_label - class of the tweets having binary values of either 0,1

We analysed the datasets for Subtask 1A, 1B, and 1C. Our analysis included visualizing the given data, which meant building word clouds of the most frequently used words, in which the biggest (the most frequently used) words are related to COVID-19. Subsequently, the frequency barplot displays the most frequently used words found after tokenizing the data, confirming our assumption. Upon further analysis, an imbalance in the label count was also observed (as seen in the label count frequency barplot).

The descriptions of the respective datasets, along with relevant visualizations are present below.

topic	tweet_id	tweet_text	class_label
COVID-19	1367208427903021	We found people who were denied COVID vaccines—even when they brought ID that fit a site's stated rules. These policies are inconsistent, sometimes arbitrary. Walgreens, Albertsons and LA now say they're working to ensure qualified people aren't rejected. https://t.co/x38ndAU62U	1
COVID-19	13704118516249354	the new federal website for covid19 vaccines coming may 1 will help users find vaccination sites not schedule vaccinations instead federal tech teams will be deployed to help states and localities with their scheduling websites per the white house covid19 response team	0

Table 1

Examples of tweets from subtask 1A.

3.1. Subtask 1A

The training, test and val datasets provided by the organizers have sizes of 2122, 149, and 195 samples respectively. The training dataset consisted of 447 check-worthy tweets and test data set had 129 check-worthy tweets. As also observed from frequency bar plot 1, the data set is highly imbalanced.

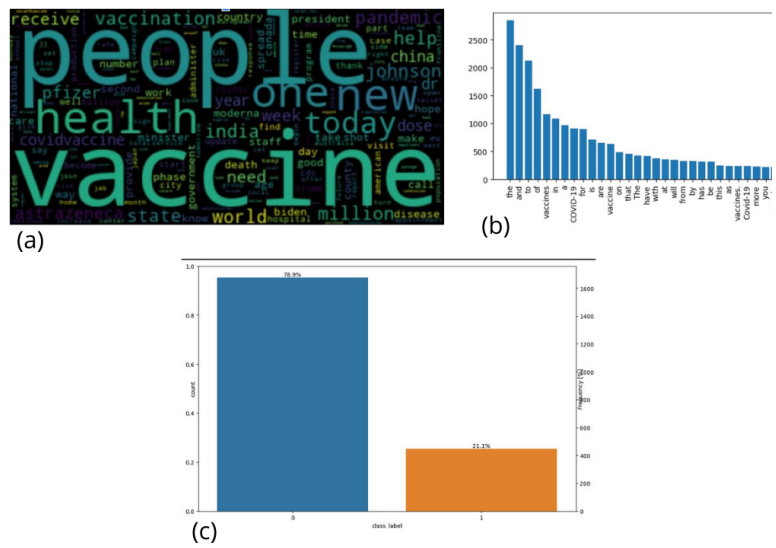


Figure 1: For Subtask 1A, English Dataset: (a) Word-cloud, (b) frequency plot validating word-cloud, (c) label distribution

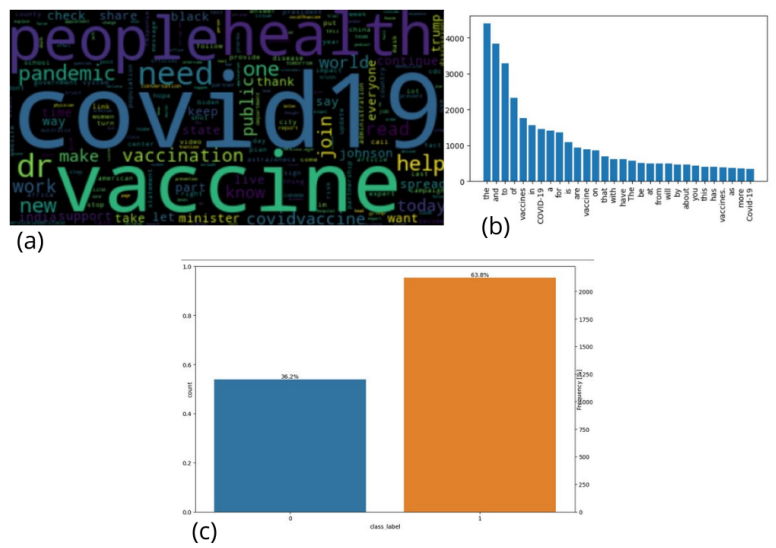


Figure 2: For Subtask 1B, English Dataset: (a) Word-cloud, (b) frequency plot validating word-cloud, (c) label distribution

3.2. Subtask 1B

The training, test and val datasets provided by the organizers have sizes of 3324, 251 and 307 respectively. The training dataset consisted of 1202 verified factual claim tweets and the test data set had 337 verified factual claim tweets. As also observed from the class_label frequency

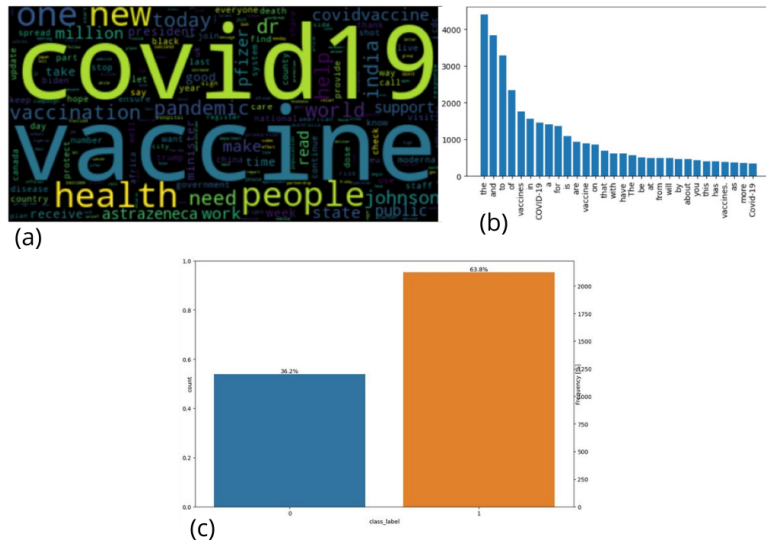


Figure 3: For Subtask 1C, English Dataset: (a) Word-cloud, (b) frequency plot validating word-cloud, (c) label distribution

bar plot 2 , the data set is slightly imbalanced, with more verified than unverified factual claims.

the latest doc	topic	tweet_id	tweet_text	class_label
COVID-19		13691451858215444e	It's an elbow bump on the arrival of the AstraZeneca vaccine. On Sunday, Minister Greg Hunt, Julia Gillard and Dr Brendan Murphy received their vaccination. They also highlighted the need to get credible #COVID19 information from sources like https://t.co/MOM9rv9OKi https://t.co/Yqz5EJPrut	1
COVID-19		13700523805579428	females have a more robust immune system that can produce more antibodies in response to vaccines may be related to hormones estrogen can cause immune cells to produce more antibodies and testosterone can suppress the prod of immune chemicals httpst:cozvqzq9hj3g	0

Table 2
Examples of tweets from subtask 1B.

3.3. Subtask 1C

The val, testing and training datasets provided by the organizers have sizes of 195, 251 and 2122 respectively. The training dataset consisted of 447 harmful tweets and test data set had 129 harmful tweets. As also observed from the class_label frequency bar plot 3 , the data set is slightly imbalanced, with more harmful tweets than non-harmful tweets.

4. System Description

4.1. Preprocessing

The text data present in the dataset needs to be processed in a manner that eliminates the non-useful components and presents the text in a way in which our model can give meaningful

topic	tweet_id	tweet_text	class_label
COVID-19	13676696127616122	florida governor ron deorsantia may be selling vaccines to high dollar donors state sen leader farmerforflsen and state ag commissioner nikkifried want the fbi to investigate after the miamiherald broke the story floridaman floridavaccine httpst coppzrwmpro httpst conqrhv80cpn	1
COVID-19	13690092112172155	Hey older Ohio #GenX, ages 50 and up are eligible for COVID-19 vaccines starting Thursday(though I don't know how long it will take to schedule a vaccine). Type 2 diabetics and those w/end-stage renal disease will also be eligible https://t.co/46xA5T8Erk	0

Table 3
Examples of tweets from subtask 1C.

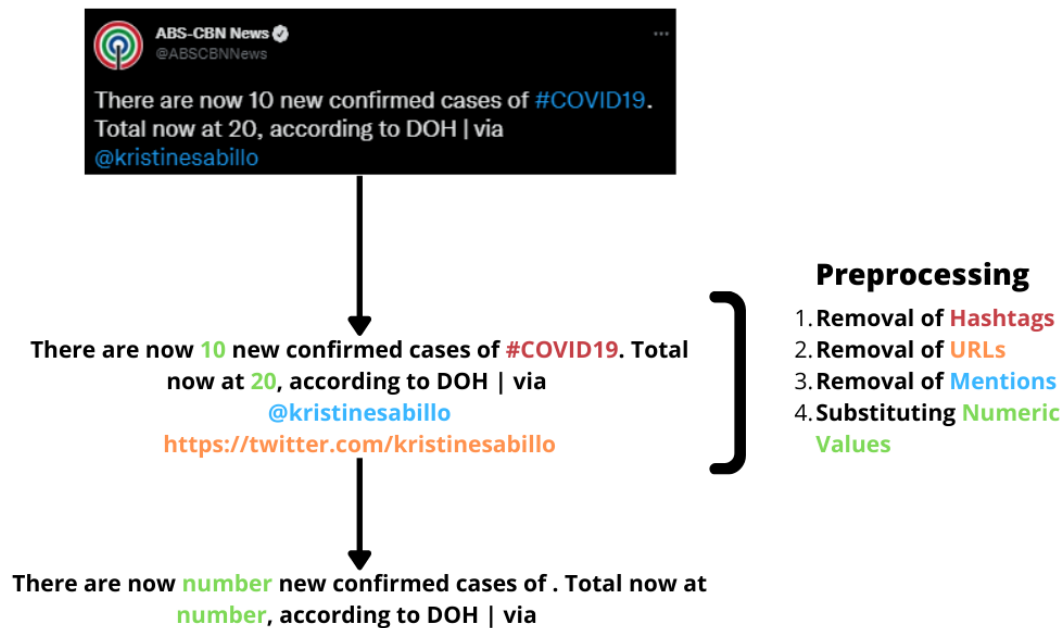


Figure 4: An example of preprocessing methodology used by our system.

results. Preprocessing the textual dataset involved the following steps:

1. Removal of hashtags:

Hashtags didn't seem to play an important role in determining the labels for the given subtasks during our initial validation runs, therefore we decided to remove them from the tweet data points in our datasets.

2. Removal of URLs:

Links or URLs don't necessarily contribute to the textual component of the tweet as they

Task	English	Bulgarian	Dutch	Total
Task 1A	2122	1871	923	4916
Task 1B	3324	2710	1950	7984
Task 1C	3323	2708	1946	7977

Table 4

This table represents the total size of our training sets following the data augmentation technique used by us.

lack semantic meaning. We removed the URLs from the tweets, however considering that the presence of URLs may play an important role in determining specific tweet attributes, we added a categorical feature to represent URLs.

3. Removal of Mentions:

We removed mentions represented by @username since they don't hold relevance to the given tweets.

4. Substituting numeric values: The presence of numerical values is an important factor in identifying characteristics such as verifiability as a number represents an objective fact that can be verified or negated. To make numeric representation uniform, we have replaced numeric figures with the string "number". Since our text is being conditioned by a large pre-trained model, we believe that it has the capability to pick up on the given contextualisation.

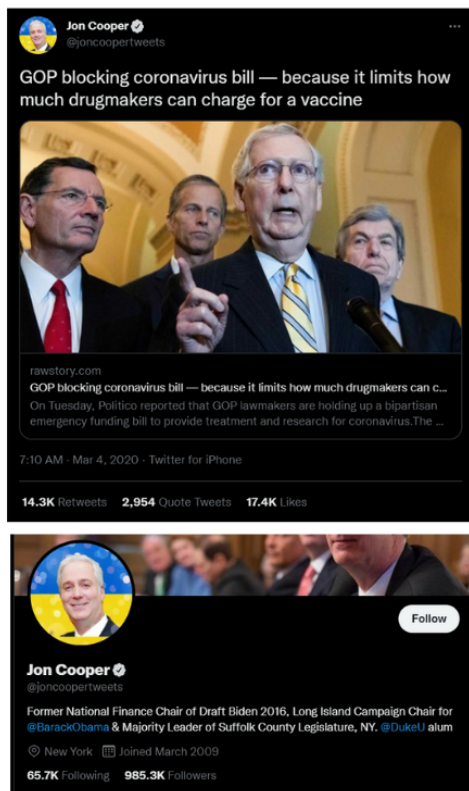
4.2. Data Augmentation

Data augmentation is the practice of increasing the diversity of training samples without collecting new data [6, 7]. We capitalised on the presence of multiple language datasets and used them to increase the quantity and diversity of training data points for Subtask 1A, Subtask 1B, and Subtask 1C.

The original training set for Subtask 1A, 1B, 1C for English consists of 2122, 3324, 3323 tweets, respectively. We translated the Bulgarian and Dutch datasets to English using the Google Translate library. We then appended the translated datasets for the respective tasks to the English dataset, yielding final datasets of sizes 4916, 7984, and 7977 for Subtasks 1A, 1B, and 1C respectively. This marked a 231%, 240%, 240% increase in the training size, respectively. Table 4 reflects the sizes of the individual training sets and the final size of the augmented sets.

4.3. Feature Extraction Through Twitter API

The purpose of this task was to identify relevant claims in tweets and to achieve this, we decided to extract additional features that help in contextualising the textual content of the tweet. [8] The additional features extracted include information about the tweet itself (the number of likes, retweets, and the presence of a URL in the tweet) as well as information about the author of the tweet (the number of followers, the number of users followed by the author, the number of previous tweets and whether the author has a verified account). Intuitively, the reason behind this is that given attributes of a tweet such as check-worthiness, verifiability, and harmful nature depend not just on the text of the tweet but the amount of engagement it has received as well as the digital footprint of the author.



Feature	Value
likes	17478
rts	14455
followers	958555
following	64763
posts	329424
verified	1
url	1

Figure 5: An example of the features extracted using the Twitter API. The tweet is from the training set of Task 1A.

As we can see in Fig 5 in reference to subtask 1A, the fact that makes the text content, has been posted by someone with strong political authority, makes it more susceptible to be check-worthy.

The features were extracted using the Twitter API. Fig 6 shows the final representation of a datapoint in our dataset. The features have been summarised below:

Numerical Features:

- followers: number of followers of the author
- following: number of users followed by the author
- posts: number of previous posts by the author
- likes: number of likes on the tweet
- rts: number of retweets of the tweet

All numerical features default to 0 when the relevant values are not available for like if the tweet has been deleted or the user has a private account.

Categorical Features:

	topic	tweet_id	tweet_url	tweet_text	class_label	followers	following	posts	likes	rts	verified	url
0	COVID-19	1.359351e+18	http://twitter.com/user/status/135935094335617...	India's gift of number COVID-19 vaccines arl...	1	49708	1155	2496	33298	7957	1	1
1	COVID-19	1.350166e+18	http://twitter.com/user/status/135016568806166...	Here's what I'm doing while I wait my turn for...	0	1288978	358	27473	14028	4005	1	1
2	COVID-19	1.369750e+18	http://twitter.com/user/status/136974953915491...	This afternoon, I'm hosting an event with the ...	0	21894707	12	3082	11936	1910	1	1
3	COVID-19	1.350165e+18	http://twitter.com/user/status/135016499568693...	Help shops like mine stay open. Mask up, avoid...	0	1288978	358	27473	7760	1868	1	1
4	COVID-19	1.370008e+18	http://twitter.com/user/status/137000807648978...	As part of the ongoing nationwide vaccination ...	1	721379	312	24220	6141	1715	1	1

Figure 6: A look at the dataframe representing the final dataset after feature extraction.

- verified: 1 if the user is verified, 0 otherwise
- has_url: 1 if the tweet contains a URL, 0 otherwise

4.4. Multimodal Model

Several methods have been used previously to combine textual input with additional features of a different modality. Models which combine transformers with visual inputs include ViLBERT [9], VLBERT [10] and MMBT [11]. Similarly, models exist which adapt to audio, visual, and text modalities such as MulT [12]. Knowledge graph embeddings [13] have also been used to combine additional textual features to transformers. The dataset we trained on this task took tabular data as input consisting numerical features and categorical features in addition to tweet texts.

Our model consists of individual Multi-Layer Perceptrons (MLPs) on the numerical and categorical features. The outputs of the MLPs are concatenated with the outputs of the transformer before the final classification layer. The method can be represented by the following equation:

$$m = x || MLP(c) || MLP(n) \quad (3)$$

where $||$ is the concatenation operator, x is output of transformer, m refers to the multimodal representation and c, n represent the categorical and numerical features respectively.

The transformer used for this task was BERT (cite). BERT refers to Bidirectional Encoder Representations (cite). It uses bidirectional transformers (cite) pretrained using a combination of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). It learns deep bidirectional representations by jointly conditioning on both left and right context layers.

Fig 7 represents the approach used by us for subtasks 1A, 1B, and 1C.

5. Experimental Setup

The models have been trained on an augmented training set consisting of the English translated training sets of Bulgarian and Dutch combined with the English dataset for each subtask. The translation was done using the *googletrans* library in python.

The models were developed using the Multimodal-Toolkit [14], which combines transformers [15] with tabular data. The transformer model chosen was *bert-base-uncased* from Hugging-Face [16]. A batch size of 8 with a variable learning rate was used to train the model. The combine method for the model is *individual_mlps_on_cat_and_numerical_feats_then_concat*

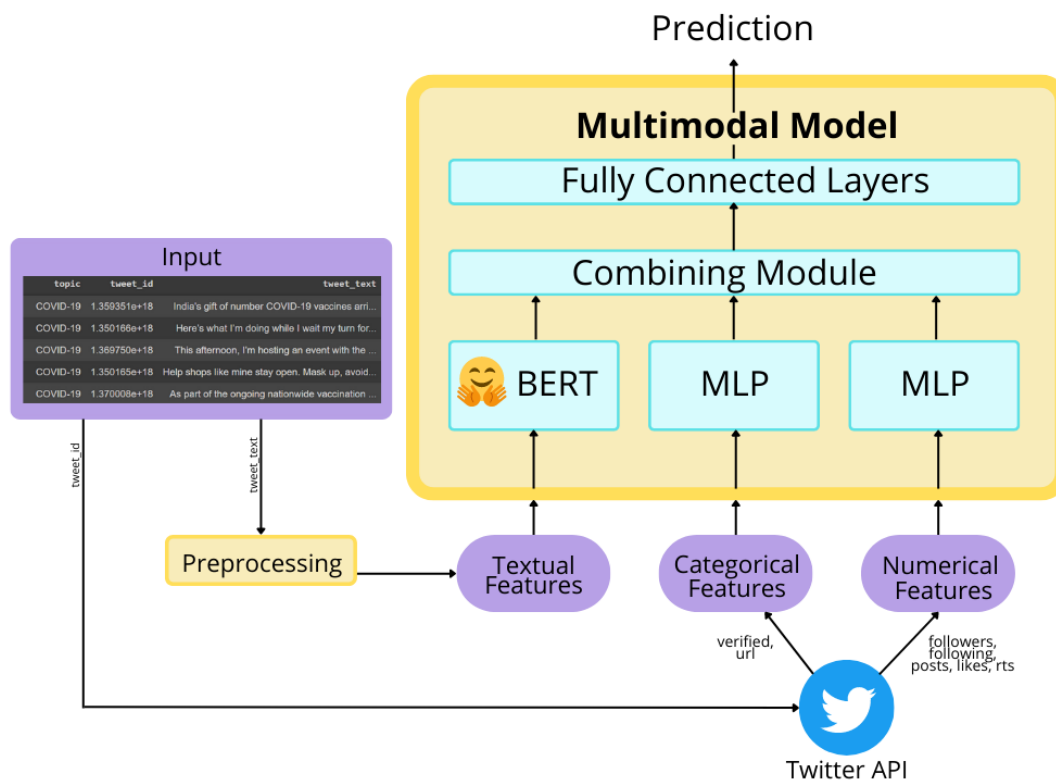


Figure 7: The framework of the multimodal model used by our system.

with a dropout ratio of 0.1 and the division ratio as 4 for the MLPs. The models were trained on Google Colab GPU for 1 epoch each.

6. Results

Our results for the three subtasks of Task 1 in which we participated are as follows:

1. Subtask 1A:

We ranked 9 out of the 13 teams that submitted for this subtask. The F1 score for the positive class for this task was 0.500.

2. Subtask 1B:

We ranked 2 out of the 10 teams that submitted for subtask 1B. The accuracy achieved on the test set by our model was 0.749. This accuracy score is very close to the score of the team ranking first on the leaderboard (0.761).

3. Subtask 1C:

We ranked 2 out of the 11 teams that submitted for subtask 1C. Our model achieved an F1 score of 0.361 on the test set, coming close to the team that ranked first having an F1 score of

Participants (userid/team-name)	F1 (positive class)
asavchev	0.698
nicuBuliga	0.667
Team_PoliMi-FlatEarthers	0.626
mkutlu	0.561
fraunhofersit_checkthat22	0.552
Team_RUB_DFL	0.525
hinokicrum	0.522
elfsong	0.519
TonyTTTTT	0.500
Asatya	0.500
Sanjana	0.482
PreronaTarannum	0.478
Team_NLP&IR@UNED	0.469
random-baseline	0.253

Table 5
Results for Subtask 1A, English on the test set.

Participants (userid/team-name)	F1 (Acc)
Team_PoliMi-FlatEarthers	0.761
Asatya	0.749
Team_NLP&IR@UNED	0.725
asavchev	0.713
nicuBuliga	0.709
Team_RUB_DFL	0.709
Sanjana	0.709
hinokicrum	0.665
mkutlu	0.641
random-baseline	0.494

Table 6
Results for Subtask 1B, English on the test set.

0.397.

For all the three subtasks, the results are on the English test set.

7. Discussion

Our model has given competitive results on all the subtasks we participated in. It performs especially well on Subtask 1B and Subtask 1C where we achieved scores of 0.749 (Accuracy) and 0.361 (F1) respectively. We ranked second on both these subtasks.

While our system gives comparable results to other teams in Subtask 1A, there is a performance gap compared to our performance in Subtask 1B and Subtask 1C. We believe that this primarily arises from a difference in the amount of training data. Our system had a uniform pipeline that performed data augmentation by combining the translated training sets of Bulgarian and Dutch. For subtask 1B and 1C, this resulted in datasets of size 7984 and 7977 tweets,

Participants (userid/team-name)	F1 (postive class)
nicuBuliga	0.397
Asatya	0.361
asavchev	0.361
Team_NLP&IR@UNED	0.347
mkutlu	0.329
ogozcelik	0.300
hinokicrum	0.281
francesco_lomonaco	0.280
Team_RUB_DFL	0.273
Team_PoliMi-FlatEarthers	0.270
random-baseline	0.200
Sanjana	0.000

Table 7

Results for Subtask 1B, English on the test set.

while for Subtask 1A, the training size was limited to 4916 tweets. Moreover, we believe, the criteria for a tweet to be check worthy requires more subjective knowledge of the tweet as compared to detection of verifiability and harmfulness. Verifiability can be determined by the presence of numbers and objective facts and harmful tweets have common patterns. Both of these attributes are not present in reference to check-worthiness of a tweet.

8. Conclusion

The task of identifying relevant claims in tweets such as check-worthiness, verifiability, and harmfulness are becoming increasingly important in a world where a significant proportion of information flow happens through the web, and the point of dissemination is often social media platforms like Twitter. Task 1 of the CheckThat! 2022 aims specifically at identifying relevant claims in tweets.

The methodology used by our team involved data augmentation by translating appropriate datasets to English, followed by feature extraction from Twitter to get relevant attributes of the tweet and the author to make a better judgement in the task of identifying relevant claims. Finally, we used a model which combined BERT with tabular features and performed a multimodal analysis of the problem set. Our model achieved competitive results in all the subtasks we participated in.

The future work for our team involves optimising our current approach. We can work with more datasets for a larger scope of data augmentation. Further, we have the opportunity to work with different transformer-based models such as DistilBERT [17], RoBERTa [18], ALBERT [19], ERNIE[20] and T5[21]. Further scope of this paper involved other concatenation techniques for tabular data with textual data such as using Multimodal Adaptation Gates (MAG) [22]. We can also work on dealing with class imbalance for improving evaluation metrics using different class weighting schemes[23].

Acknowledgments

We would also like to thank the organisers of CheckThat! Lab 2022 for conducting this shared task.

References

- [1] M. D. Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, W. Quattrociocchi, The spreading of misinformation online, *Proceedings of the National Academy of Sciences* 113 (????) 554–559. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1517441113>. doi:10.1073/pnas.1517441113.
- [2] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), *Advances in Information Retrieval*, Springer International Publishing, Cham, 2022, pp. 416–428.
- [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [4] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, M. Wiegand, M. Siegel, J. Köhler, Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection, in: *Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022*, Bologna, Italy, 2022.
- [5] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022*, Bologna, Italy, 2022.
- [6] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. H. Hovy, A survey of data augmentation approaches for NLP, *CoRR* abs/2105.03075 (2021). URL: <https://arxiv.org/abs/2105.03075>. arXiv:2105.03075.
- [7] P. Liu, X. Wang, C. Xiang, W. Meng, A survey of text data augmentation, in: *2020 International Conference on Computer Communication and Network Security (CCNS)*, 2020, pp. 191–195. doi:10.1109/CCNS50731.2020.00049.
- [8] R. Ahuja, A. Chug, S. Kohli, S. Gupta, P. Ahuja, The impact of features extraction on the sentiment analysis, *Procedia Computer Science* 152 (2019) 341–348. URL: <https://>

www.sciencedirect.com/science/article/pii/S1877050919306593. doi:<https://doi.org/10.1016/j.procs.2019.05.008>, international Conference on Pervasive Computing Advances and Applications- PerCAA 2019.

- [9] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, CoRR abs/1908.02265 (2019). URL: <http://arxiv.org/abs/1908.02265>. arXiv:1908.02265.
- [10] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: pre-training of generic visual-linguistic representations, CoRR abs/1908.08530 (2019). URL: <http://arxiv.org/abs/1908.08530>. arXiv:1908.08530.
- [11] D. Kiela, S. Bhooshan, H. Firooz, D. Testuggine, Supervised multimodal bitransformers for classifying images and text, arXiv preprint arXiv:1909.02950 (2019).
- [12] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569. URL: <https://aclanthology.org/P19-1656>. doi:10.18653/v1/P19-1656.
- [13] M. Ostendorff, P. Bourgonje, M. Berger, J. M. Schneider, G. Rehm, B. Gipp, Enriching BERT with knowledge graph embeddings for document classification, CoRR abs/1909.08402 (2019). URL: <http://arxiv.org/abs/1909.08402>. arXiv:1909.08402.
- [14] K. Gu, A. Budhkar, A package for learning on tabular and text data with transformers, in: Proceedings of the Third Workshop on Multimodal Artificial Intelligence, Association for Computational Linguistics, Mexico City, Mexico, 2021, pp. 69–73. URL: <https://www.aclweb.org/anthology/2021.maiworkshop-1.10>. doi:10.18653/v1/2021.maiworkshop-1.10.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: <http://arxiv.org/abs/1706.03762>. arXiv:1706.03762.
- [16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface’s transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019). URL: <http://arxiv.org/abs/1910.03771>. arXiv:1910.03771.
- [17] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: <http://arxiv.org/abs/1910.01108>. arXiv:1910.01108.
- [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [19] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, CoRR abs/1909.11942 (2019). URL: <http://arxiv.org/abs/1909.11942>. arXiv:1909.11942.
- [20] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, ERNIE: enhanced language representation with informative entities, CoRR abs/1905.07129 (2019). URL: <http://arxiv.org/abs/1905.07129>. arXiv:1905.07129.
- [21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, CoRR

- abs/1910.10683 (2019). URL: <http://arxiv.org/abs/1910.10683>. arXiv:1910.10683.
- [22] W. Rahman, M. K. Hasan, A. Zadeh, L. Morency, M. E. Hoque, M-BERT: injecting multimodal information in the BERT structure, CoRR abs/1908.05787 (2019). URL: <http://arxiv.org/abs/1908.05787>. arXiv:1908.05787.
- [23] M. Suri, PiCkLe at SemEval-2022 Task 4: Boosting Pre-trained Language Models with Task Specific Metadata and Cost Sensitive Learning, in: The 16th International Workshop on Semantic Evaluation, 2022.