FOREAL: RoBERTa Model for Fake News Detection based on Emotions

Vladislav Kolev, Gerhard Weiss[a] and Gerasimos Spanakis[b]
Department of Data Science and Knowledge Engineering, Maastricht University, Paul-Henri Spaaklaan 1, Maastricht, The Netherlands

Keywords: Natural Language Processing, Fake News, Emotion Classification, Emotion Analysis, Sentiment Analysis, Transformers, RoBERTa.

Abstract: Detecting false information in the form of fake news has become a bigger challenge than anticipated. There are multiple promising ways of approaching such a problem, ranging from source-based detection, linguistic feature extraction, and sentiment analysis of articles. While analyzing the sentiment of text has produced some promising results, this paper explores a rather more fine-grained strategy of classifying news as fake or real, based solely on the emotion profile of an article's title. A RoBERTa model was first trained to perform Emotion Classification, achieving test accuracy of about 90%. Six basic emotions were used for the task, based on the prominent psychologist Paul Ekman - fear, joy, anger, sadness, disgust and surprise. A seventh emotional category was also added to represent neutral text. Model performance was also validated by comparing classification results to other state-of-the-art models, developed by other groups. The model was then used to make inference on the emotion profile of news titles, returning a probability vector, which describes the emotion that the title conveys. Having the emotion probability vectors for each article's title, another Binary Random Forest classifier model was trained to evaluate news as either fake or real, based solely on their emotion profile. The model achieved up to 88% accuracy on the Kaggle Fake and Real News Dataset, showing there is a connection present between the emotion profile of news titles and if the article is fake or real.

1 INTRODUCTION

It does not come as a surprise that the spread of false information is one of the major issues in our society nowadays. Fake news have a single "first and foremost" goal - gain attention. This is usually done by using a specific type of trigger keywords or phrasing that aims at invoking a strong emotional response in the reader. As a result, fake news may be more "emotionally saturated" or simply have a different emotion profile, compared to real news. Analyzing the emotion saturation per text may enable one to determine exactly what patterns of emotion are mostly associated with fake news. Assuming most real news aim at informing the general public in a somewhat neutral manner, fake news would rely on more emotionally saturated content, which would trigger a limbic response in the reader. Based on this assumption, a technique for identifying news as fake or at least suspicious could be developed.

There is already research on Sentiment Analysis of fake news which shows promising results. Doing Emotional Analysis is taking things one step further and obtaining more granular results. Using state-of-the-art techniques in Natural Language Processing (NLP), an Emotion Classification RoBERTa model was built, which was trained to classify a piece of text as one of the following emotions: happiness, anger, disgust, fear, sadness, surprise, neutral. The first six "basic emotions" are based on the work of the psychologist Paul Ekman and his taxonomy (Ekman, 1992). Additionally, neutral was added as a seventh emotion category. The model was trained on text annotated with emotions - it learned to classify text, based on the dominant emotion. Furthermore, the numeric probability output of all seven emotions per text was also taken into account and put into an "emotion probability vector" - such a vector represents the emotional arousal of the text, which would allow for an emotion profile to be created. The model was then used on news for emotion classification. Finally, a Random Forest Binary Classifier was used for fake and real news recognition, based on the emotion probability vector values per article's title.

[a] https://orcid.org/0000-0002-6190-2513
[b] https://orcid.org/0000-0002-0799-0241

Kolev, V., Weiss, G. and Spanakis, G.
FOREAL: RoBERTa Model for Fake News Detection based on Emotions.
DOI: 10.5220/0010873900003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 2*, pages 429-440
ISBN: 978-989-758-547-0; ISSN: 2184-433X

Summarizing it into steps, it looks as follows:

1. Combine data from datasets into one balanced dataset, which would match short corpus of text to one of the six basic emotions by Paul Ekman and the neutral category/emotion in addition.
2. Fine-tune and test RoBERTa base model from the HuggingFace library on this dataset for classification and generate a probability distribution over all emotions per instance of text
3. Use already fine-tuned RoBERTa for inference over the Fake and Real News Dataset - this means associating the title of each article with its dominant emotion and its emotion probability distribution
4. Train a Binary Random Forest classifier using the emotion probability distribution data that was generated per article as independent variables and the fake/real label as a dependent variable
5. Perform Experiments with different emotions and compare prediction accuracy and F1 score results

Three main **Research Questions (RQ)** were explored in this paper:

- **RQ1** - How well does a RoBERTa model perform Emotion Classification, compared to other methods used by other research groups?
- **RQ2** - Do fake news have a different emotion profile compared to real news and is such potential difference statistically significant?
- **RQ3** - Can Emotion Classification of news text be used to improve fake news detection?

2 RELATED WORK

There are multiple techniques developed to detect fake news, explored by numerous groups. For example, detecting rumors in microblog posts using propagation structure via kernel learning (Ma et al., 2017), early detection of fake news on social media through propagation path classification with recurrent and convolutional networks (Liu and Wu, 2018) and fake news detection on twitter using propagation structures (Meyers et al., 2020). Furthermore, other groups like (Zhang and Ghorbani, 2020) have done extensive analysis of fake news as a phenomenon and reviewed existing techniques to identify them. (Shu et al., 2017) explored fake news detection on social media specifically. The team conducted a survey where they present a comprehensive review of detecting fake news on social media, including fake news characterizations on psychology and social theories, existing algorithms from a data mining perspective, evaluation metrics and representative datasets. This includes an extensive problem definition, possible Feature Extraction and model construction techniques. Their motivation is to facilitate further research into this area, since social media became an incubator for the spread of misinformation.

(Pérez-Rosas et al., 2017) tested an automatic fake news detection approach, using an SVM classifier and five-fold cross-validation, with accuracy, precision, recall, and F1 measures. The team managed to achieve results that are comparable to human ability to spot fake content, producing over 90% accuracy for fake news recognition in domains such as politics and technology.

(Ruchansky et al., 2017) devised a more general and complex model called CSI (Capture, Response, Score) that takes into account the text, the response an article receives and the users who source it. The first module, Capture, captures the abstract temporal behavior of user encounters with articles, as well as temporal textual and user features, to measure response as well as the text. The second component, Score, estimates a source suspiciousness score for every user, which is then combined with the first module by Integrate to produce a predicted label for each article. The team achieved an accuracy of 89.2% for a Twitter dataset and 95.3% for the Weibo dataset.

In the field of Affective Computing, both Sentiment analysis and Emotion Analysis are crucial. Due to the more complex nature of the latter problem, more work has been done on Sentiment analysis in the past. This is starting to change. There already are multiple projects and research groups focused on identifying the underlying emotions within a corpus of text. However, not much research has been done on Emotion Analysis of Text with respect to fake news recognition, profiling and classification.

(Demszky et al., 2020) created GoEmotions - one of the largest fine-grained and manually annotated datasets, labeled for 27 emotions. Furthermore, they tested a BERT transformer model on it, achieving a macro-averaged F1 score of 0.46 % for all 27 emotions and macro-averaged F1 score of 0.64 % using Ekman's emotion taxonomy, which is also used in this paper. A baseline RoBERTa model was built for the sake of comparing results to the BERT model of Demszky et al. This is described in sections below.

(Ghanem et al., 2020) argued that the emotion profile of set of fake news (propaganda, hoax, clickbait, and satire) would be different, compared to real news. The team used a set of different emotions to evaluate a piece of text, one of which is the set of six

emotions also used in this paper - namely Paul Ekman's proposed six basic emotions. They proposed an emotionally-infused (EIN) LSTM neural network to detect fake news on Twitter and in news articles. The model utilizes both emotional features and text features to classify a piece of text as one of the following categories: propaganda, hoax, clickbait, satire and real news. The model achieved up to 81% accuracy detecting on news Articles, 65% on Twitter posts and up to 96 % accuracy on a dataset focused specifically on clickbait (Stop_Clickbait). They compared their EIN model against different baselines, showing that emotionally-infusing an LSTM model does produce better overall results. They also tested a separate model, which was trained only on emotional features (similar to the model being proposed in this paper). They achieved up to 50% and 52% accuracy on news articles and Twitter posts respectively.

(Paschen, 2019) analyzed the differences in the emotional framing of the message content in fake and real news. What their group found is that there is a significant difference between the overall emotional sentiment portrayed in the titles and that fake news are substantially more negative with regard to the emotion dimensions disgust and anger compared to real news articles. Furthermore, real news were found to be more "joyful" than fake news. What they also showed is that article titles are a good differentiator on emotions between Fake and real news. These results coincide with some of the results for this paper, even though different datasets were used for training.

3 DATASETS

3.1 Emotion Classification Dataset

Data from four datasets was randomized and sampled to create one balanced dataset for the RoBERTa Emotion Classification. Each emotion is represented with about 2000 samples, forming the final dataset with 14 958 samples overall. What all datasets have in common is that each contains a small piece of text, no more than 100 characters, where each instance is annotated with the dominant emotion describing it. The main dataset used is the HuggingFace Emotion Dataset by (Saravia et al., 2018), which consists of English Twitter messages. It contains text, associated with the following 6 emotions - fear, joy, sadness, anger, surprise and love. As already mentioned, this paper focuses on the 6 basic emotions, described by the psychologist Paul Ekman, therefore only the first five emotions were taken into account. Additionally, the combined data collected from the other three datasets is associated with the disgust emotion, neutral emotion as well as the surprise emotion. The reason behind this is because disgust is missing from the HuggingFace dataset in the first place and also there were not enough instances for surprise either. Therefore, data from the ISEAR (International Survey on Emotion Antecedents and Reactions) (Scherer and Wallbott, 1990), DailyDialogue (Li et al., 2017) and Emotion Stimulus (Ghazi et al., 2015) datasets was collected in order to have a balanced dataset with all 6 basic emotions plus the neutral emotion. Please refer to Figure 4 for an overview of the Emotion Distribution of the final combined dataset.

3.2 Fake and Real News Dataset

The Fake and Real News Dataset, created by (Ahmed et al., 2018), was used for training and testing the Random Forest classifier. The dataset consists of articles and contains information about the Title, Text, Subject and Date, where each instance is labeled as either 0 (Fake) or 1 (Real). Real news were collected from Reuters.com to represent truthful information. Fake news were collected from Kaggle.com - the articles stem from websites that Politifact (a fact-checking organization in the USA) identified as unreliable and spreading false information. Table 1 shows some common statistics for this dataset, as well as for the Emotion Classification Dataset.

Table 1: Dataset Statistics.

	Emotion Classification Dataset	Fake and Real News Dataset
Training Size	11966	35918
Validation Size	1496	-
Test Size	1496	8980
Avg token length	20	21
Balanced	Yes	Yes

4 METHODS

This paper is primarily focused on a model that was named FOREAL - Fake Or Real Emotion Analyzer. The model is a combination of a RoBERTa model that performs Emotion Analysis and a Binary Random Forest classifier model that classifies articles as either fake or real, based on the emotion analysis done by RoBERTa. Figure 2 represents the model diagram. The Methods section describes the different components of the model architecture and the other baseline models used to perform experiments.

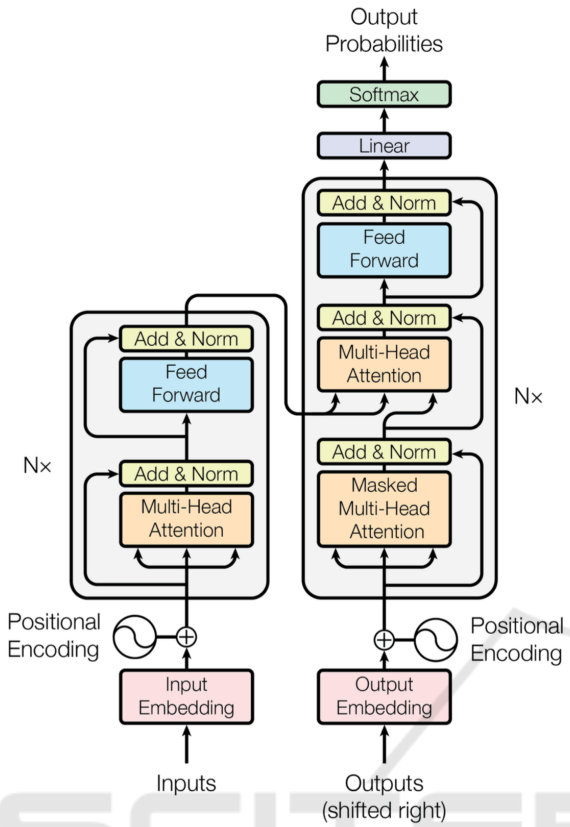

Figure 1: Transformer model architecture ((Vaswani et al., 2017)).

Since the RoBERTa model used for this paper is heavily based on the BERT model, a short general description of how BERT and RoBERTa work algorithmically is added in the next two sub-sections.

4.1 The BERT (Bidirectional Encoder Representations from Transformers) Model

The initial BERT model was proposed by (Devlin et al., 2018) and is a language representation model that relies on pre-training and fine-tuning a Transformer model (Fig. 1). Transformers utilize the concept of self-attention, originally described by (Vaswani et al., 2017), which takes input sequence and decides at each step which other parts of the sequence are important. BERT model architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in (Vaswani et al., 2017) (2017) and released in the tensor2tensor library.

BERT takes two concatenated sequences as input, where each sequence consists of more than one text sentence. Both segments are taken at once as input and are delimited with special tokens - [CLS] Sequence 1 [SEP] Sequence 2 [EOS]. Furthermore, concepts of Masked Language Model (MLM) and Next Sentence Prediction (NSP) are used during pre-training. MLM is a process of randomly masking some percentage of the input tokens, and then predicting those masked tokens. NSP on the other hand is about understanding the relationship between two sentences, which is not directly captured by language modeling. It is a binary task that predicts whether or not a sentence B follows sentence A. This means that each training example consists of a text pair (segment A, segment B). The starting point is always the first segment A. In 50% of the cases, the second segment B is the actual segment that follows segment A. While in the other 50% of the cases, BERT randomly selects a segment from the whole corpus. Going in-depth into the mechanics and workings of BERT and Transformers is not the aim of this paper.

BERT uses a Byte-Pair Encoding approach and so does the RoBERTa model. What this means, as the name suggests, is that encoding is done on a byte level, instead of on unicode characters. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

4.2 RoBERTa (Robustly Optimized BERT Approach) Model

RoBERTa stands for Robustly Optimized BERT approach and is a model proposed by (Liu et al., 2019). Both BERT and RoBERTa models rely on the Transformer model architecture, Fig. 1 as a basis, with slight differences in pre-training and fine-tuning. Compared to the BERT model, the team behind RoBERTa had three major improvements over the BERT model which go as follows: RoBERTa is trained with dynamic masking , FULL-SENTENCES without NSP (Next-Sentence Prediction) loss, large mini-batches and a larger byte-level BPE (Byte-Pair Encoding). Furthermore, RoBERTa is pre-trained on more data, longer sequences and with bigger batch sizes. Other than that, the RoBERTa base model resembles the BERT base model architecture of 12 encoder layers, 768 hidden and embedding size and 12 attention heads.

Embedding for RoBERTa is also the same as for BERT, however, the RoBERTa vocabulary is much larger and therefore the model is utilizing more parameters

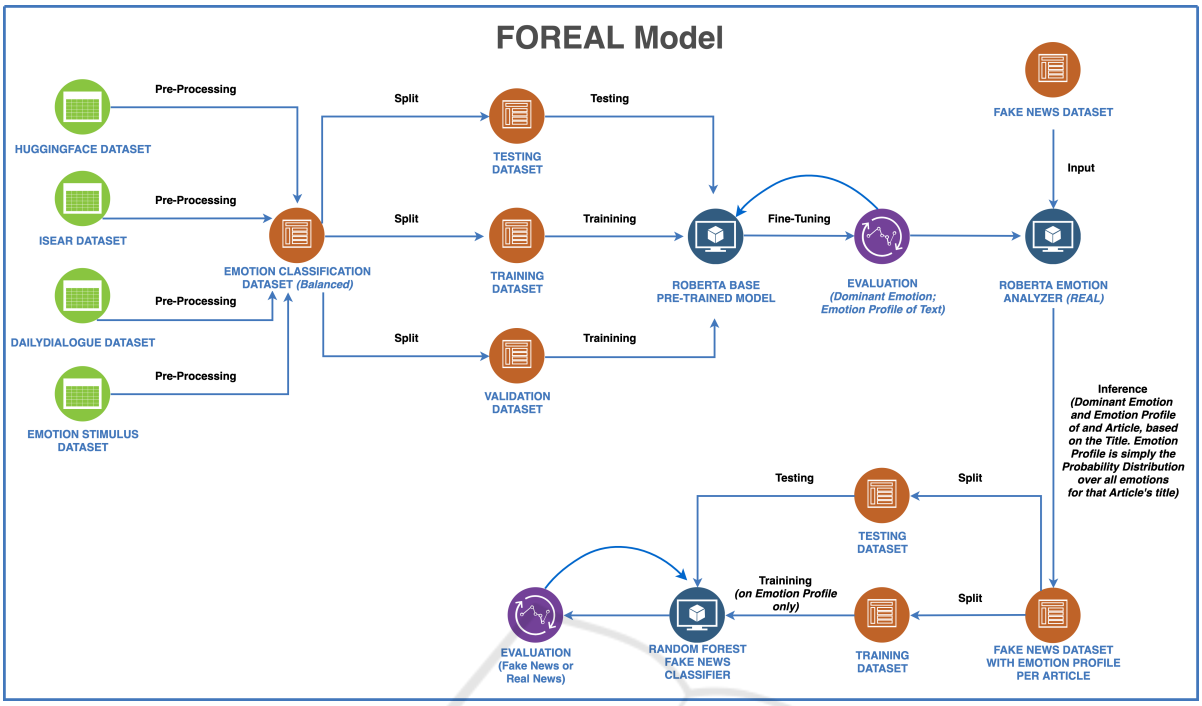

Figure 2: Diagram of how the REAL and Random Forest models work to form the entire FOREAL model.

4.3 Fake or Real Emotion Analyzer (FOREAL)

The FOREAL model is simply the combination of a RoBERTa Emotion Analyzer (REAL) and a Random Forest Fake News Classifier working together. Please refer to Fig. 2 for a visual depiction of how the entire FOREAL model works. The following two subsections explain how those models work:

4.3.1 RoBERTa Model for Emotion Analysis (REAL)

The RoBERTa base model used for this paper was fine-tuned on a combination of datasets, which had a block of text classified as one of the following six basic emotions - joy, anger, fear, sadness, surprise or disgust. Another "neutral" emotional category was added in order to represent text which is not emotionally infused. Data from all datasets was randomized and combined into the Emotion Classification Dataset.

What was used is a "roberta-base" model from the HuggingFace library, which is already pre-trained and could be fine-tuned for the Emotion Classification task. Even though the model is initially trained on cased text, all experiments done in this paper are with non-capitalized text due to the nature of the training data. Obtaining cased training data of better quality could also be an important factor for future improvements of the model. Gradient clipping technique was added to avoid having the gradients exploding. Dropout layer was added for Regularization sake and a fully-connected output layer with a SoftMax function so that a probabilistic output of all emotion classes would be produced. For more information on the hyperparameters, please see Table 2

The REAL model was trained over 10 epochs, where the model with both lowest validation loss and highest validation accuracy was chosen in the end. 20% of the data was used for testing, where this data was further split into 10% for the validation set and 10% for the test set. Once the emotion classifier model was trained and tested on the data, the same model was used for inference on the Fake and Real News Dataset. The emotion classifier evaluates all articles to return an emotion probability vector, which represents the probability of each emotion, based on the news title - this would also be referred to as emotion profile in this paper. The highest probability value is chosen in order to associate news with a single dominant emotion as well. Next step is applying a Binary Classifier on the dataset with fake news, which now also contains the emotion data per instance - this process is described in the next section.

Table 2: Hyperparameters for both Pretraining and Finetuning the RoBERTa-base model.

Hyperparam	Pretraining	Finetuning
Dropout	0.1	0.3
Gradient Clipping	0.0	1.0
Batch Size	8k	16
Weight Decay	0.01	0.01
Learning Rate Decay	Linear	Linear
Warmup Steps	24k	0
Learning Rate	-	2e-5
Max Epochs	-	10
Peak Learning Rate	6e-4	-
Number of Layers	12	-
Hidden size	768	-
FFN inner hidden size	3072	-
Attention heads	12	-
Attention head size	64	-
Attention Dropout	0.1	-
Max Steps	500k	-
Adam ε	1e-6	-
Adam β_1	0.9	-
Adam β_2	0.98	-

4.3.2 Binary Classification - Random Forest

A Random Forest algorithm was used for Binary Classification with 100 estimators and using "gini impurity" approach for measuring the quality of a split. Random Forest is a supervised learning algorithm, which builds numerous Decision Trees by using Bootstrap Aggregating techniques, also called Bagging, developed by (Breiman, 1996). Furthermore, 10-fold Cross-Validation was applied for evaluation purposes primarily. Even though Random Forest already applies the concept of Bagging, Cross-Validation is also useful in the sense that Bagging does sampling with replacement, whereas Cross-Validation is doing splits on the data at hand only. The model is trained on the emotion profile probability values for each article's title and classifies it as either fake or real. 20% of the data was used for testing.

Multiple other Binary Classifiers were tested and Random Forest was chosen because of the better overall accuracy and F1 results.

A typical Random Forest algorithm was used, which builds many individual Decision Trees during training. For the fake vs real news binary classification, the process could be described as follows (adapted from (Ronaghan, 2018)):

$$\underbrace{C}_{\text{Gini Impurity}} = \underbrace{\sum_{i=1}^{N} f_i(1-f_i)}_{\substack{f_i \text{ is the frequency of label } i \text{ at a node and } N \text{ is the number of unique labels}}} \tag{1}$$

$$\underbrace{ni_j}_{\text{Importance of Node } j} = \underbrace{w_j C_j}_{\substack{\text{Weighted number of samples } w \text{ times impurity } C \text{ for node } j}} - \underbrace{w_{left(j)} C_{left(j)}}_{\substack{\text{child node from left split on node } j}} - \underbrace{w_{right(j)} C_{right(j)}}_{\substack{\text{child node from right split on node } j}} \tag{2}$$

$$\underbrace{fi_i}_{\substack{\text{Importance of feature } i}} = \frac{\sum_{j:node\,j\,splits\,on\,feature\,i} ni_j}{\sum_{k \in all nodes} ni_k} \tag{3}$$

$$\underbrace{normfi_i}_{\text{Normalization}} = \frac{fi_i}{\sum_{j \in all features} fi_j} \tag{4}$$

$$\underbrace{RFfi_i}_{\substack{\text{Importance of feature } i \text{ calculated from all trees}}} = \frac{\sum_{j \in all trees} normfi_i j}{T} \tag{5}$$

4.4 Baseline Models

4.4.1 Baseline Fake or Real Sentiment Analyzer (FORSAL)

A second RoBERTa model was built for Sentiment Analysis (FORSAL), classifying articles as either negative, neutral or positive. This model was used as a baseline for comparing the results of the emotion classifier model to the sentiment classifier. Data from the same datasets was used for this model, having collected samples for all 7 motions (joy, fear, sadness, anger, disgust, surprise, neutral). Emotions were converted to sentiment as follows:

1. anger, fear, disgust, sadness - Negative
2. neutral, surprise - Neutral
3. joy - Positive

Furthermore, the resulted dataset was balanced, having the same size of approximately 2000 instances per sentiment, as for the Emotion Classification task. Other than that, the same hyperparameters were used for this model, as for the emotion classifier RoBERTa model.

4.4.2 REAL Model Trained on GoEmotions and ISEAR

The REAL model was fine-tuned on two other datasets for validation purposes. The model was trained on the GoEmotions dataset by (Demszky et al., 2020), using Ekman's emotion taxonomy. It was also trained on the ISEAR dataset separately. Results were then compared to the results obtained by (Demszky et al., 2020) in order to check how well the RoBERTa model performs Emotion Classification, compared to other groups and approaches (BERT).

4.4.3 Baseline Fake News RoBERTa (FNR)

Finally, a baseline FNR model was used, which was trained and tested directly on the Fake and Real News Dataset. This model does not implement any specific emotional aspect, like all other models. The FNR model was directly trained on the news titles text data. Same hyperparameters were used for this model, as for the emotion classifier RoBERTa model.

5 EXPERIMENTS

Multiple experiments were performed with both the REAL model for Emotion Classification and the FOREAL model for fake news detection and profiling. A 10-fold Cross-Validation was performed for the underlying Random Forest Classifier model only - this means that all FOREAL results are averaged over 10 iterations. Cross-Validation was not performed for any of the RoBERTa models in any of the experiments due to time constraints related to fine-tuning.

Furthermore, performance was compared to different baseline models and approaches by other groups. Accuracy, F1 score for Binary Classification and macro-averaged F1 score for Multiclass Classification were used as evaluation metrics:

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \quad (6)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

$$F1_{macro} = \frac{\sum_{i=1}^{N} F1_i}{\underbrace{N}_{\text{Number of Classes}}} \quad (8)$$

The experiments can be described as follows:

5.1 Baseline BERT and REAL Model Performance Comparison

Emotion Classification performance of REAL was compared to the BERT model by the group of (Demszky et al., 2020). The models were trained and separately tested on both the ISEAR and the GoEmotions (by (Demszky et al., 2020)) datasets.

5.2 REAL Emotion Classification Experiments

As described above, the REAL model was fine-tuned on different sets of emotions, in order to compare performance. This was done in order to determine if adding emotions and making classification more fine-grained would also improve performance. Four experiments were performed:

- Four emotions - anger, fear, joy, sadness
- Five emotions - anger, fear, joy, sadness, disgust
- Six emotions - anger, fear, joy, sadness, surprise, disgust
- Seven emotions - anger, fear, joy, sadness, surprise, disgust, neutral

5.3 FOREAL Fake/Real News Classification Experiments

The performance of the FOREAL model was tested in terms of classifying news as either fake or real. Same as for the previous experiment with the REAL model, 4 experiments were performed with different sets of emotions in order to check if a more granular approach would improve fake/real news classification. First the Fake and Real News Dataset was emotionally assessed - this means that each article's title was processed by FOREAL and its emotion profile was inferred, based on the sets of emotions described in the previous sub-section. Thus an emotion profile per article was induced. Furthermore, fake/real news classification was performed, based on this emotion profile data alone.

5.4 Compare Baseline FORSAL and FOREAL Performance

The FORSAL model was compared to the FOREAL model in order to determine if Emotion Analysis would improve fake news recognition, compared to Sentiment Analysis. Both underlying RoBERTa models were trained with the same hyperparameters and on the same data. The underlying Random Forest

Classifier was cross-validated, using 10-fold cross-validation for both the FORSAL and FOREAL models.

5.5 Compare Baseline FNR and FOREAL Performance

The baseline FNR model was compared to the FOREAL model in order to determine how well the Emotional Analysis approach would perform, compared to a RoBERTa model that is trained on the text title data directly. Both underlying RoBERTa models were trained with the same hyperparameters, however the FNR model was trained over 3 epochs only, since more training was not needed.

5.6 Statistically Significant Difference between Fake News Emotions and Real News Emotions

A non-parametric Mann-Whitney U Test was done for each of the seven emotions across real and fake news - the reason behind this is that the sample data per emotion does not follow a normal distribution and sample size may vary. Furthermore, Chi-Square Test of Independence was considered and ruled out as a possible test to be performed, due to the input percentage form of the data. As per (McHugh, 2013), the data in the cells should be frequencies or counts of cases rather than percentages or some other transformation of the data.

6 RESULTS

6.1 Baseline BERT Compared to REAL

Table 3 shows the macro-averaged F1 score for the baseline BERT model, compared to the REAL model on both ISEAR and GoEmotions datasets. As could be observed, RoBERTa achieved slightly better results on the ISEAR dataset (0.74) and slightly worse on GoEmotions (0.59), compared to BERT (0.70 and 0.64 respectively).

6.2 REAL Emotion Classification Results

The REAL Emotion Classification results for different sets of emotions achieved results around the 90% mark, which is a good indicator of how well the model connects emotions to text. Adding more emotions does not necessarily improve or worsen the Emotion Classification per se, nonetheless RoBERTa does a good overall job of classifying emotions. The results can be found in Table 4.

Table 3: Baseline comparison between REAL model and the Demszky et al BERT model, tested on both GoEmotions and ISEAR datasets. The macro-averaged F1 score was measured.

Model	F1 GoEmotions	F1 ISEAR
REAL	59%	**74%**
BERT (Demszky et al)	64%	70%

Table 4: REAL Test Set accuracy on classifying emotions.

Model	Accuracy	F1
4 Emotions	93.6%	94%
5 Emotions	95.9%	96%
6 Emotions	91.5%	91%
7 Emotions	89.04%	89%

6.3 FOREAL Model - Emotion Classification and Fake News Detection

Testing the whole FOREAL model on a different set of emotions shows that the more granular the approach is i.e. more emotions to describe the text are included, the better the result is. This goes for both classification accuracy and the F1 metric, which is a combination of precision and recall. Results vary from 70.1% accuracy for a model with only 4 emotions to about 87.5% for a model with 7 emotions. Since both standard deviation values for accuracy and F1 are small, what can be concluded is that the results are closely spread to the mean. Please refer to Table 5 for results of the experiments.

6.4 FOREAL and FORSAL for Fake News Detection

Comparing the FOREAL (Emotion Analyzer) to the baseline FORSAL (Sentiment Analyzer), Table 6 shows that the FOREAL is better at classifying Fake and real news, achieving accuracy of about 87.5%, while the FORSAL model achieved an accuracy of 79.4%.

6.5 FOREAL and Baseline FNR for Fake News Detection

Finally, the baseline FNR model was tested against the FOREAL model on classifying fake and real

Table 5: FOREAL Test Set Results on Classifying Fake and Real News.

	Averaged		SD	
Model	Accuracy	Macro F1	Accuracy	Macro F1
4 Emotions	70.01%	67.6%	0.64%	0.45%
5 Emotions	74.0%	67.8%	0.38%	0.67%
6 Emotions	84.9%	84.8%	0.47%	0.57%
7 Emotions	**87.5%**	**86.8%**	**0.36%**	**0.34%**

Table 6: FOREAL and FORSAL test set results on classifying Fake and real news. This is a baseline comparison, showing a more granular approach with different emotions produces better accuracy, compared to sentiment analysis alone.

	Averaged		SD	
Model	Accuracy	Macro F1	Accuracy	Macro F1
FORSAL (baseline)	80.4%	79.04%	0.88%	0.84%
FOREAL (7 Emotions)	**87.5%**	**86.8%**	**0.36 %**	**0.34%**

news. FNR achieved impressive accuracy of about 99.9%, while the FOREAL model achieved an accuracy of about 88% - clearly the FNR model performs much better than the FOREAL model. However, it is difficult to interpret the reason behind such high percentage accuracy, making it a black-box model. Furthermore, such high accuracy percentage is very suspicious, hinting at possible overfitting. This is why the model was tested on a separate fake news dataset, Getting Real about Fake News, (Risdal, 2016), which was obtained from Kaggle. Not surprisingly, FNR achieved no more than 55.2% accuracy, which clearly points to overfitting. On the other hand, having the FOREAL model trained only on emotion data clearly shows results, based solely on emotion profiles of text. Results in Table 7 are for a single sample code execution and were not cross-validated due to time constraints.

Table 7: Baseline comparison between FNR and FOREAL. Clearly FNR has much higher accuracy, but it was shown it is indeed overfitting, when comparing results to another fake news dataset. Furthermore, it is difficult to explain based on what has the FNR learned to do the classification. Results shown in the table are not cross-validated due to time and computational complexity limitations for fine-tuning the FNR model.

Model	Accuracy	F1
FNR (baseline)	**99.9%**	**100%**
FOREAL (7 Emotions)	87.5%	86.8%

6.6 Statistically Significant Difference of Emotions across Real and Fake News

Results for all emotions produced a p value way below the σ value of 0.05, essentially approaching 0. Therefore the difference for each emotion across real news instances and fake news instances is statistically significant. What this means is that for example the emotion disgust is expressed statistically different across all fake news, compared to real news.

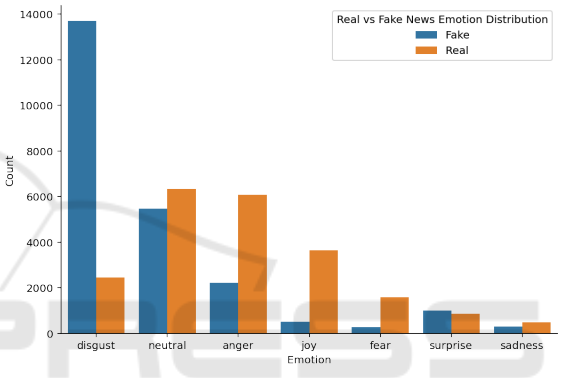

Figure 3: Comparing both emotion distributions for Real and Fake Articles, as inferred by the REAL model.

Furthermore, Fig. 3 shows the emotion distributions for real news and fake news respectively. Here only the dominant emotion is taken into account. What could be observed is that the majority of fake news are associated with the disgust emotion, whereas real news are associated with neutral, anger.

6.7 Limitations

There are certain aspects about how some of the experiments were performed, which can be considered limitations for the results obtained. One such limiting factor was the time constraint related to experiments done with the RoBERTa models - as mentioned above, none of the RoBERTa results were cross-validated, because the process of fine-tuning is very time-consuming due to computational complexity. While those experiments were indeed performed multiple times, producing almost identical results, not

a proper 10-fold cross-validation was done.

Additionally, another limitation that should be considered has to do with the experiments done comparing different number of emotions for both emotion classification and fake news classification. They were performed with specific sets of emotions, namely:

- Four emotions - anger, fear, joy, sadness
- Five emotions - anger, fear, joy, sadness, disgust
- Six emotions - anger, fear, joy, sadness, surprise, disgust
- Seven emotions - anger, fear, joy, sadness, surprise, disgust, neutral

Randomizing and using different set of emotions for each category would probably produce different results (for example having surprise, disgust, anger and fear for the "Four emotions" scenario). Due to time constraints again, only the aforementioned scenarios were tested.

7 DISCUSSION AND CONCLUSION

Emotion Classification done with the REAL model achieves promising results, maintaining about a 90% accuracy on the sets of emotions that were tested - from 4 to 7 emotions. Furthermore, compared to results by other groups, such as (Demszky et al., 2020), REAL performs either as well as other state-of-the-art models or even better (Table 3). This points to the robustness of the REAL model as an Emotion Classifier, therefore answering **RQ1**.

Additionally, fake news have shown to have a different emotion profile for the dataset used, compared to real news. Most of fake news are associated with disgust and are less emotionally versatile, whereas real news were shown to be more neutral, angry but also joyful (Figure 3).The significance of this difference was statistically confirmed by performing the Mann-Whitney U Test, which answers **RQ2**.

Furthermore, fake news classification results for the FOREAL model show that every time an emotion is added, both accuracy and F1 increase. This indicates that emotions and their intensities can serve as a good predictor for fake news detection.

In addition, the experiment done with the FORSAL baseline model and the FOREAL model shows that Emotion Analysis improves the task of fake/real news classification, compared to Sentiment Analysis when it comes to both.

While the results of the FNR fake news Classification model may seem to be undermining the performance of the FOREAL model at first, one might argue this is actually not the case. As mentioned in the previous section, the FNR was tested on a completely separate fake news dataset, showing 55.2% accuracy. Also, FNR is more of a black box, where the choices and reasons behind the classifications are not easy to understand and explain. This could also lead to a biased model or a model that is overfitted on the data and the noise present in the dataset. On the other hand, the FOREAL model is solely trained to classify real and fake news, based on emotion intensities, which makes it easier to explain the decisions it is making.

Last but not least, the emotion profile of real news is much different and more versatile, compared to fake news. The dominant emotions in real news seem to be neutral and anger, whereas for fake news it is mainly disgust. This also makes sense from a "real-world" perspective, since it is expected that real news are more neutral and not mainly associated with negative emotions - for example, big part of the real news are associated with joy as well (about 4000 instances). However, this could not be said about fake news with less than 400 joy instances. This also matches the results by Paschen et al, described in the Related Work section.

What all those results point to is a strong connection between the emotion profile of an article's title and if this article is considered fake or real news (**RQ3**).

7.1 Future Work

Needless to say, a lot more research is needed in order explore the connection between the emotion of text and whether or not this piece of text is considered fake news. For example, the work here was done with the title of the news article, whereas the actual body of the article can be explored as well in further research. Additionally, different sets of emotions from different datasets can be used as a predictor, which are not necessarily based on Paul Ekman's taxonomy. Another aspect would be to test different state-of-the-art NLP models for emotion analysis (instead of RoBERTa) and compare performance. Different transformer, LSTM neural network or Convolutional Neural Network (CNN) models can be explored for this purpose. Apart from this notion, the field would greatly benefit from newer and more robust datasets with respect to fake news. Much expertise and effort is required to identify and distinguish between fake and real news, even more so to make a robust dataset collection of such. However, building a better "lie detector" for news may depend on it.

REFERENCES

Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer.

Ahmed, H., Traore, I., and Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Cortiz, D. (2021). Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. *arXiv preprint arXiv:2104.02041*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Ekman, P. (1992). Are there basic emotions?

Ghanem, B., Rosso, P., and Rangel, F. (2020). An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–18.

Ghazi, D., Inkpen, D., and Szpakowicz, S. (2015). Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Govi, P. (2020). Classify emotions in text with bert.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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

Liu, Y. and Wu, Y.-F. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-second AAAI conference on artificial intelligence*.

Ma, J., Gao, W., and Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

Meyers, M., Weiss, G., and Spanakis, G. (2020). Fake news detection on twitter using propagation structures. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, pages 138–158. Springer.

Paschen, J. (2019). Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Risdal, M. (2016). Getting real about fake news.

Ronaghan, S. (2018). The mathematics of decision trees, random forest and feature importance in scikit-learn and spark. *Toward Data Science url*.

Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Scherer, K. and Wallbott, H. (1990). International survey on emotion antecedents and reactions (isear).

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Valkov, V. (2020). Sentiment analysis with bert and transformers by hugging face using pytorch and python.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Zhang, X. and Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.

APPENDIX

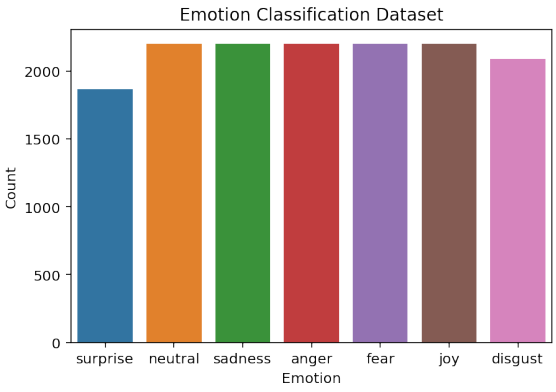

Figure 4: Distribution of emotions in the Emotion Dataset. The dataset is a combination of the Huggingface Emotion Dataset, ISEAR, DailyDialogue and the Emotion Stimulus datasets.

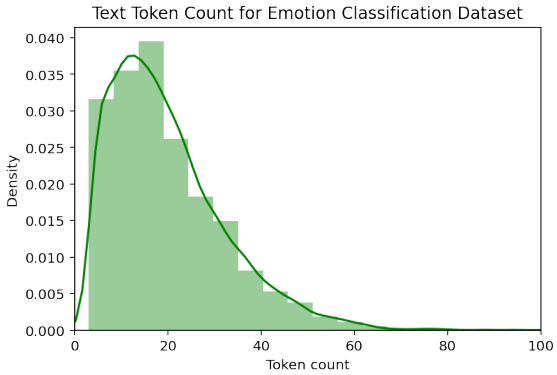

Figure 5: Majority of the Emotions Dataset data consists of text that is less than 100 tokens long.

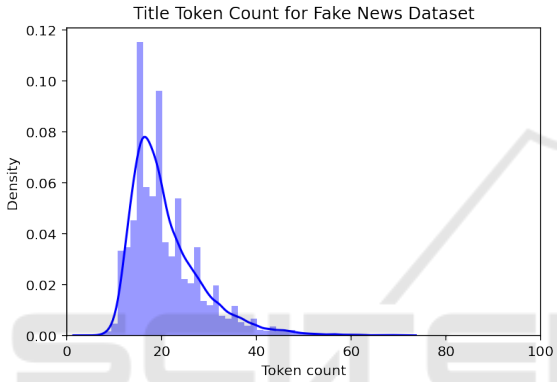

Figure 6: Same as for the Emotion Dataset, the Fake/real news titles are less than 100 tokens long.