ISSN 1870-4069

Multimodal Misinformation Detection from YouTube Videos Employing on Early and Late Fusion

Luz Elisa Gahona-Castillejos, Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad

Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico

{luz.gahona, ariel, fmartine}@inaoep.mx

Abstract. Misinformation has gained importance recently as online platforms and social networks have become so popular and influential in daily life, such as elections, public health, and the economy. Misinformation poses a challenge, and current approaches have not yet produced an effective solution. In this paper, we propose to focus on detecting multimodal misinformation in YouTube videos. Our proposal involves extracting features from text and audio independently and then combining them. Traditional features such as bag-of-words and embeddings were used to represent the text extracted from videos, and we also explored pre-trained transformations. Regarding the audio extracted from videos, a windowed feature extraction method was proposed, which includes zero crossings, root mean square (RMS), and eGeMAPS features. The best features of both modalities were combined in an early fusion, and then late fusion was used to improve misinformation detection accuracy, as shown in our experiments.

Keywords: Bag of words, transformes, RMS, models, fusion late, misinformation.

1 Introduction

The term "disinformation" refers to information that is both false and intentionally disseminated to mislead and harm others. This harmful intent is a crucial aspect of disinformation, yet it is often overlooked. In contrast, "misinformation" also involves spreading false information without the intent to cause harm [2].

In recent years, the spread of fake news and harmful language on various online platforms, especially social media, has become a significant concern. Detecting and curbing the dissemination of such content has become a priority for researchers and governments [3].

Misinformation detection is also related to fact-checking since both tasks aim to assess the veracity of claims [6]. Depending on where the information is

sourced. It may include text only, or audio only in the case of spoken interviews. Multimodal data is also common when sourced from videos, as it can consist of audio, and text when the video is transcribed.

The speech signal is complex as it carries information about the message content, speaker, language, and emotions. While many speech-processing systems work well with neutral speech in controlled environments, they often struggle with emotional speech due to the complexity of modeling and describing spontaneous speech.

The meaning of a spoken text can change based on intonation and context. For instance, "OKAY" in English can express emotions and attitudes like admiration, disbelief, consent, disinterest, or affirmation.

Therefore, understanding the text alone cannot fully capture a spoken statement's meaning [8]. The problem of misinformation has been addressed in different ways, one of them being during the COVID-19 pandemic, when a lot of misinformation was spread on YouTube and other social networks. [12] proposes a simple methodology based on NLP that can help fact-checkers detect misinformation about COVID-19 on YouTube, this work focuses on the comments of the videos (a single modality) from which they extract features that can be used for misinformation detection and creates a multi-label classifier based on transfer learning that can detect misinformative comments.

Accurate detection of misinformation content is essential to mitigate its impact and promote a more reliable and truthful information environment. By combining text and audio modalities, the characteristics of what is said and how it is being said could help in detecting misinformation more effectively.

Following this idea, in this paper, we propose a multimodal approach for detecting misinformation based on early and late fusion.

The paper is structured as follows. Section 2 reviews the related works closest to our proposal. Section 3 presents our proposal for detecting multimodal misinformation, detailing how to handle both modalities through early and late fusion. Section 4 shows the experiments carried out to validate our proposal. Finally, Section 5 includes our conclusions and future work.

2 Related Work

Previous research has explored the capabilities and limitations of NLP in identifying incorrect information. In [11], the authors propose focusing on generalization, and uncertainty and leveraging recent language models such as RoBERTa-large and GPT-4. On the other hand, [13] suggests a Multimodal Co-Attention Network (MCAN) that better fuses textual and visual features in fake news detection.

Another approach was proposed by [12], which uses YouTube video comments and pre-trained transfer learning models to generate a multi-label classifier that can categorize conspiratorial content. Today, one of the most used platforms is YouTube, where you can find a lot of misinformation about, health problems such as being overweight. Because YouTube cannot inform the user whether the

information in a video is true or not, there is a need to develop methods to recognize, from the videos, whether the information is correct.

Most misinformation detection work focuses on the text modality, with data extracted from platforms such as Twitter.

We propose from YouTube videos, taking only the audio and transcribing, this detection can be carried out.

Although the following works do not focus on the detecting of misinformation, they do focus on the recognition of emotions in a multimodal way using audio and its transcription, offering an idea of how to address misinformation detection. In [5] uses BERT, which is known to be very effective for many text classification tasks, and Mel coefficients for audio characterization. Different approaches have been proposed. For example, in the multimodal system developed by Robinet [7], the Inception-ResNet-V2 neural network is used for emotion recognition in speech. This network receives the audio's melodic frequency cepstral coefficients (MFCC) as input and RoBERTa is used to extract features from the text modality. The fusion of modalities is carried out through a closed attention mechanism.

On the other hand, in [9], the extraction of text features is addressed using the GloVe and Elmo embedding models. For audio, a convolutional neural network (CNN) is used along with a long-term recurrent memory (LSTM) model.

The features extracted from both modalities are concatenated for further processing. [7] and [9] works utilize the melodic frequency cepstral coefficients (MFCC) as the primary audio characteristic, obtained by segmenting the audio into windows. These coefficients serve as input for the respective neural networks. Another appears feature used in voice-related work is the root mean square (RMS).

Different types of fusion combine two or more modalities depending on the stage they are combined. Early fusion, also known as data-level fusion or feature-level fusion, occurs at a very early stage of model development, before any network layer. It appears at the raw or preprocessed data stage, or simply when features have been extracted from the raw data. Late fusion, also known as decision-level fusion, involves first developing a complete model for each modality. The individual models' outcomes (decisions or probabilities) are then integrated [3].

3 Proposed Approach

As can be seen, there are currently no existing methods for detecting multimodal misinformation. Therefore, we propose a new approach for multimodal misinformation detection in this paper, based on early and late fusion.

As seen in Figure 1, the transcriptions are extracted for the text modality, and the audio from the video must also be extracted to process the voice signal; both modalities are used to detect misinformation from YouTube videos. In the following subsections, we describe each stage shown in Figure 1.

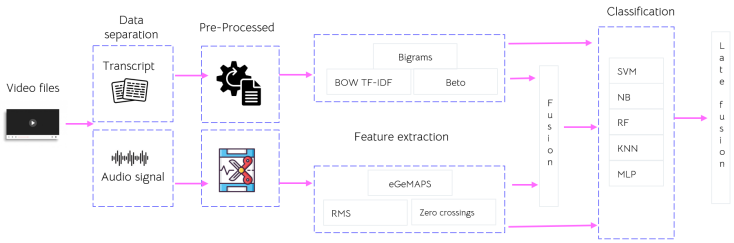


Fig. 1. Diagram of our proposal for Multimodal Misinformation Detection Employing Early and Late Fusion.

3.1 Data Separation

During this stage, we automatically obtained video transcripts to make it easier to analyze the spoken content. We also extracted the audio from the original video so that we could focus on the voice signal. This extraction enabled us to conduct a more detailed analysis of both the text and the audio.

3.2 Pre-processing

For the transcriptions, a cleaning process was carried out that included eliminating periods, commas, stop words, special characters, question marks, and exclamation marks; accents were left. Additionally, all text was converted to lowercase to ensure uniformity in the analysis. For the audio, some of them contained music, so a cleaning process was carried out to remove the music and leave only the voice. Subsequently, the audios were segmented into one-minute intervals. In cases where the audios were less than one minute, they were repeated until that duration was reached. The first minute of all the audios was taken for one set of experiments and, for another set, the minute from the moment speech was detected was used; this was done to the audios with and without music. Hamming windowing is applied, which avoids discontinuities introduced by analyzing only a fraction of the signal. The Hamming window function also has the property that its value varies smoothly from zero at the ends to one in the center, which helps avoid loss of information due to overlap between windows.

The Hamming windowing was applied to the audio with a size of $50ms$ per window and overlaps of $10ms$. The audios were also reduced to only the first 60 seconds; the rest of the audio was discarded from the analysis. If an audio did not reach that length, the signal was repeated until all had the same length, in normalize the sample size.

3.3 Feature Extraction

In this stage, the features of the video transcripts were obtained, which consisted of:

1. Bag of Words: This approach represent a document as a vector where each dimension corresponds to a single word in a predefined vocabulary. For weighing the words, we use the TF-IDF (Term Frequency-Inverse Document Frequency) by considering not only the frequency of words in the document, and their relative importance in the entire corpus.
2. N-grams: are contiguous sequences of n elements, which can be characters, words, or tokens. N-gram extraction allows for capturing contextual information by considering not only individual words, but also combinations of words that appear together. In our proposal, we use word bigrams (N=2).
3. BETO (Bidirectional Encoder Representations from Transformers for Spanish): BETO is an adaptation of BERT designed explicitly for the Spanish language. Like BERT, it uses a bidirectional transformer architecture to pre-train a language model on a large corpus of Spanish text.

The extraction of features in the audio part is for divided into two approaches. In the first, together with preprocessing, 13 Mel Frequency Cepstral Coefficients (MFCC) were extracted, where cepstral coefficients are a specific type of coefficients derived from cepstral analysis applied to a time window of the speech signal. This technique efficiently separates the two main components of information in the speech signal: excitation and vocal tract [10]. By zero crossings were obtained for each window, indicating how many times the speech signal crosses the zero level during a given period. This measure provides an overview of how the signal's energy is distributed [4]. Finally, the root mean square energy was extracted, where the energy of a signal refers to the sum of all the magnitudes of the signal. In the context of audio signals, this measure indicates the intensity signal's in terms of its volume or sound level.

The other approach consisted of using the Opensmile eGeMAPS feature set that consists of 88 static acoustic features resulting from calculating of various features on low-level descriptor functions, such as volume frequency loudness, alpha ratio, harmonic difference, mffc 1-4, loudness peak rate, among others [1].

3.4 Early Fusion

Once the individual features from each modality have been extracted, they are concatenated. For this experiment, we combined the features obtained from the text's term Frequency-Inverse Document Frequency (Tf-IDF) representation of the text with the individual features extracted from the audio signals. By integrating these features, we aim to leverage the strengths of both modalities: the semantic richness captured by Tf-IDF from the textual data and the emotional cues embedded in the audio features. This multimodal approach enhances the overall representation of the data, providing a more comprehensive basis for misinformation detection. The concatenation process involves aligning the feature vectors and creating a unified input for the classification model, thereby enabling the model to learn from text and audio characteristics simultaneously.

3.5 Classification

We trained several classifiers to detect misinformation the classifiers used were:
Linear SVM (Support Vector Machine): This classifier aims to find a linear hyperplane that separates the data into different classes in a multidimensional space. It is particularly beneficial when the data is linearly separable. By maximizing the margin between the classes, SVM ensures robust classification even with high-dimensional data.
Naive Bayes (NB): This classifer is based on the Bayes' theorem. It calculates the probability that a given instance belongs to a particular class based on the likelihood of observing certain features. Naive Bayes is known for its simplicity, efficiency, and effectiveness, especially with large datasets, and when the assumption of feature independence (naivety) holds approximately true.
Multilayer Perceptron (MLP): This is a basic form of artificial neural networks. MLPs are flexible and can learn complex relationships in the data through hidden layers. They are particularly effective in capturing non-linear patterns and interactions among features.
Random Forest (RF): This ensemble learning method is known for its performance on large datasets with many features. It is robust to missing data and outliers and does not require extensive data preparation. Random Forest builds multiple decision trees and merges them to get a more accurate and stable prediction.
K-Nearest Neighbors (KNN): This is a simple and effective instance-based learning classifier. KNN does not make any assumptions about the underlying data distribution and can capture complex relationships in the data. It classifies instances based on the majority label among the nearest neighbors, for our experiments we used K= 3.

3.6 Late Fusion

In addition to feature-level fusion, we implemented a late fusion strategy using five different classifiers: Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP); and Random Forest (RF). Each classifier independently predicts the emotion class, and the final classification is determined by majority voting among the classifiers. This late fusion approach aims to improve the robustness and accuracy of the classification by combining the strengths of various classification algorithms.

4 Experiments

In the experiments, the data set was initially divided into 70% for training the models and 30% for testing. The IDs of both the audio and text elements were maintained in the same order to allow for proper analysis.

4.1 Dataset

Misinformation detection in the domain of Obesity and Overweight in Spanish language (MOOSP) is a dataset on obesity and overweight misinformation in Spanish, where queries such as "treatment for obesity and overweight", "risk factors for obesity", and prevention of obesity and overweight" were used to search for videos of interest. The database has a total of 243 videos, of which 113 are misinformation and 130 are informative, therefore the data collection is not imbalanced. The average length of the videos is 5.08 minutes, the most extended video in the database is 12.41 minutes, and the shortest one is 40.50 seconds.

The Language Technologies Research Group (LABTL) of Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) developed the dataset used in this research. We thank Doctors Manuel Montes y Gómez, Luis Villaseñor Pineda, and Master Jennifer Pérez Santiago for providing access to these data, which were fundamental for the advancement of this study.

4.2 Text Experiment

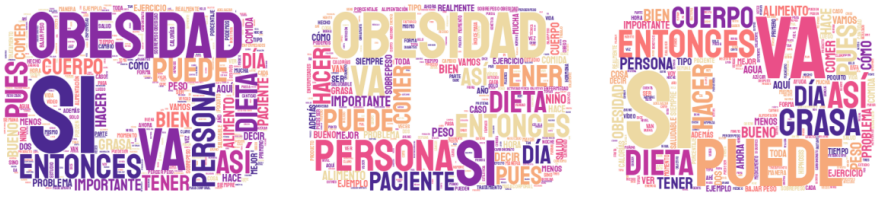
After obtaining the feature vectors, five different classifiers were used for the text modality. Each classifier processed the vectors individually to analyze the textual information. The majority voting combined the results of the five classifiers to ensure the precision and robustness of the predictions; this approach allowed for a final decision based on most individual predictions.

4.3 Audio Experiments

In the case of audio, several experiments were carried out to evaluate different scenarios. First, we worked with audio that contained music. From these audios, features from the first minute were extracted, and vectors were generated and then processed using the five classifiers mentioned above. This experiment was repeated by taking the audio sample from when speech was detected instead of the first minute. Subsequently, the same experiments were carried out with audio from which the music had been removed, leaving only the voice. In all cases, after the five classifiers analyzed the feature vectors, a majority voting was again applied to combine the results. In this way, it was possible to evaluate the influence on the performance and compare the results obtained with and without the presence of music in the audio.

4.4 Text Results

Once the corpus was pre-processed, the word cloud was obtained, which allows for identifying the most frequent words in a set of text quickly and easily. Each word is represented in the cloud with a size proportional to its frequency of appearance in the text, which can be helpful to obtain a quick understanding of the main themes or keywords present in the text; in the word cloud of the whole corpus you can see the filler words or statements such as "si", "va", This



(a) Word cloud of the corpus (b) Word cloud of the informative class (c) Word cloud of the misinformation class

Fig. 2. Word cloud of the corpus and the classes of informative and misinformation.

Table 1. Classification results (in terms of accuracy) with the different classifiers on the TF-IDF representation.

Classifer	Accuracy
SVM	0.55
MLP	0.55
Naive Bayes	0.58
Random Forest	0.49
KNN	0.52
Majority Voting	0.52

Table 2. Classification results (in terms of accuracy) with the different classifiers on the bigrams representation.

Classifer	Accuracy
SVM	0.79
MLP	0.78
Naive Bayes	0.78
Random Forest	0.74
KNN	0.71
Majority Voting	0.86

is seen in the Figure 2(a) while in the informative class words that stand out are "obesidad", "persona", and "dieta", see figure 2(b) as well as a more formal language, while in the word cloud of the misinformation class, more fillers are observed, the word "grasa" 2(c).

When reviewing the results in Tables 1-3 with the different types of features used, we can see that the representation of texts through Bigrams got the best accuracies. Neither TF-IDF nor Beto representations outperformed one of the results in Table 2.

Table 3. Classification results (in terms of accuracy) with the different classifiers on the BETO´s vector representation.

Classifer	Accuracy
SVM	0.27
MLP	0.52
Naive Bayes	0.68
Random Forest	0.40
KNN	0.51
Majority Voting	0.43

Table 4. Classification accuracy using eGeMaps for different classifiers without music.

Classifer	Accuracy
SVM	0.27
MLP	0.52
Naive Bayes	0.63
Random Forest	0.38
KNN	0.40
Majority Voting	0.39

4.5 Audio Results

After applying the methodology to the audio, the results without music are shown in Tables 4 and 5. In these tables, it was observed that the eGeMaps feature set was the most favorable. This finding suggests that the features extracted from eGeMaps provided relevant and discriminative information for classifying the audio into the desired classes. This result supports the effectiveness of eGeMaps as a set of features for detecting misinformation in audio.

For the audio data, as shown in Tables 4 and 5, the results obtained with the individual features were not very favorable for this modality. These tables included the audio without music, starting the analysis when the dialogue begins, as this approach provided better results compared to other representations, such as preserving the music or starting from the beginning of the videos.

4.6 Fusion

Using the results of the classifiers of the modalities that had the best results and then applying a majority vote, we obtain what is shown in the tables 6, which reflects an improvement compared to the unimodal result and their respective majority voting.

From the above experiments, we can appreciate that early feature concatenation effectively improved data representation. Subsequently, using majority voting in late fusion allowed for the outperformance of the individual modalities, reflected in a significant improvement in accuracy.

Table 5. Classification accuracy for different classifiers using RMS features without music.

Classifer	Accuracy
SVM	0.27
MLP	0.32
Naive Bayes	0.64
Random Forest	0.30
KNN	0.42
Majority Voting	0.34

Table 6. Classification accuracy of majority vote using eGeMaps and RMS with Bigrams.

Features	Classifer	Accuracy
EgeMaps + Bigrams	Majority Vote	0.85
RMS + Bigrams	Majority Vote	0.78

5 Conclusions

Detecting misinformation poses significant challenges depending on the analyzed data modality being analyzed. While textual features such as bigrams effectively capture nuances in text-based messages, acoustic features from eGeMaps emerge as a good option for misinformation analysis in audio formats. Integrating these modalities through concatenation does not necessarily improve misinformation detection accuracy, highlighting the complexities and potential redundancies introduced.

The results revealed that early feature fusion was particularly effective, providing a robust and significantly improved model performance. Subsequently, by applying late fusion, the individual modalities' performance surpassed the individual modalities' misinformation detection. From our experiments, We can conclude that our proposal captured contextual information, but also improved classification accuracy compared to models using only one modality. These findings underscore the importance of integrating multiple modalities and applying early and late fusion techniques to improve misinformation detection.

For future work, it is promising to explore using pre-trained language models, such as BERT or RoBERTa, for text feature extraction. These models have proven highly effective in various natural language processing tasks because they capture complex contexts and semantic relationships within text. Integrating these advanced features could significantly improve misinformation detection systems.

On the other hand, the application of transformers in audio characterization also deserves attention. Wav2Vec and HuBERT have shown great potential in extracting relevant features and improving tasks such as speech recognition, emotion detection, and acoustic event classification. By taking advantage of

transformers' capabilities, a richer and more accurate representation of audio signals can be obtained, which could lead to better results in various applications.

We are also considering exploring the possibility of adding the sequence of images from the video as another modality to analyze body language.

Acknowledgments. This research was partially supported by the National Council of Humanities, Sciences, Technologies, and Innovation of Mexico (CONA-HCyT) through its graduate study scholarship program.

References

1. Opensmile python documentation, https://audeering.github.io/opensmile-python/
2. Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G.D.S., Shaar, S., Firooz, H., Nakov, P.: A survey on multimodal disinformation detection. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 6625–6643. International Committee on Computational Linguistics (2022)
3. Ayetiran, E.F., Özgöbek, O.: A review of deep learning techniques for multimodal fake news and harmful languages detection. IEEE Access 12, 76133–76153 (2024)
4. Bleda, S., Francés, M., Marini, A., Martínez, C.: Herramientas software para la docencia de la señal de voz en ingeniería técnica de telecomunicaciones (2024), https://ice.ua.es/es/jornadas-redes-2012/documentos/posters/246141.pdf
5. Das, M., Raj, R., Saha, P., Mathew, B., Gupta, M., Mukherjee, A.: Hatemm: A multi-modal dataset for hate video classification. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 17, pp. 1014–1023 (2023)
6. Hossain, T., IV, R.L.L., Ugarte, A., Matsubara, Y., Young, S., Singh, S.: Covidlies: Detecting covid-19 misinformation on social media. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics (2020)
7. Khurana, Y., Gupta, S., Sathyaraj, R., Raja, S.: Robinnet: A multimodal speech emotion recognition system with speaker recognition for social interactions. IEEE Transactions on Computational Social Systems (2022)
8. Koolagudi, S., Rao, K.: Emotion recognition from speech: a review. International Journal of Speech Technology 15(2), 99–117 (2012)
9. Koromilas, P., Giannakopoulos, T.: Deep multimodal emotion recognition on human speech: A review. Applied Sciences 11(17), 7962 (2021)
10. Laynez, D.B.: Sistemas de Verificación Automática de Locutor, Capítulo 3. Proyecto fin de carrera, ingeniería superior de telecomunicaciones, Universidad de Sevilla, Sevilla, España (2012)
11. Pelrine, K., Imouza, A., Thibault, C., Reksoprodjo, M., Gupta, C., Christoph, J., Godbout, J.F., Rabbany, R.: Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6399–6429 (2023)
12. Serrano, J.C.M., Papakyriakopoulos, O., Hegelich, S.: Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In: Proceedings of the 1st Workshop on NLP for COVID-19. Association for Computational Linguistics (2020)

13. Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z.: Multimodal fusion with co-attention networks for fake news detection. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. pp. 2560–2569 (2021)