UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
ELECTRONICS ENGINEERING DEPARTMENT
VALPARAISO - CHILE

CONTRADICTION DETECTION WITH MACHINE LEARNING

ALINE CHRISTINA COVARRUBIAS GARCIA

as partial fulfillment of the requirements for the professional diploma of
CIVIL ELECTRONICS ENGINEER

Advisor	:	WERNER CREIXELL F. Ph.D.
Co-Advisor	:	MAURICIO ARAYA L. Ph.D.

AUGUST 2023

ACKNOWLEDGEMENTS

I would like to first thank my siblings Maite, Fernanda, and Thiago for being one of my great motivations to grow and be better every day. I hope that one day you will be able to do what makes you happy and feel fulfilled. Also to my family.

I would like to thank the professors of this university who have answered my million questions with much patience. With their good disposition and dedication they helped me learn and advance further.

To my mentor Professor Werner for his honest advice, which helped me to remember that I can look for a job that inspires me and not work out of obligation. Also, his great motivation and enthusiasm generate a cordial and collaborative atmosphere among all, where we can all feel comfortable.

Finally to all my friends from life and from university whom I can trust and rely on. Because two heads are better than one and having a good time in everyday life is also important.

Abstract

In general, unreliable information is present in our daily lives and it becomes a challenge for technology to understand this phenomenon. Being able to detect misinformation and "Fake News" is the first step to address this problem and Natural Language Processing (NLP) is the field of Machine Learning that can be up to the task. As contradictions can be a source of misinformation, the objective of this work is to implement a machine learning system that identifies the relationship between two sentences to detect if there is a contradiction. The methodology is to apply transfer learning to the BETO model to detect contradictions in Spanish. Accuracy over 90% is obtained by the model and adding new datasets can improve performance for a specific context of contradictions.

Contents

List of Tables

List of Figures

1 | Introduction

We live in what is called the second information revolution. Digital technologies allow us to interact with each other quickly, regardless of our location. We can access large amounts of information, and everybody can publish content. This technology significantly impacts society, making access to knowledge, like valid academic data and relevant news possible. However, it also has its downside; social media can also represent a threat, like the spread of disinformation and “Fake news” (1). This can happen intentionally or not, but it can negatively impact society’s well-being (2).

As we can perceive, during the COVID-19 Pandemic, social media was full of people talking about the outbreak, and a large amount of information spread worldwide. The information on social media can be based on scientific facts and reliable sources, but it can also be misleading. This became a significant concern for the World Health Organization because fake news spread baseless health advice, anxiety, and racism, harming efforts to overcome the pandemic (3; 4). The World Health Organization (WHO) worked with Facebook, TikTok, and YouTube, among other platforms, to address this issue. For example, an alert message to the user can appear onscreen over COVID-19-related videos on Instagram, urging the observer to search for reliable sources of information [3, 4]. Also, direct links to WHO’s web page would appear when searching for “covid” on Facebook, Pinterest, and Instagram, as figure 1.1 shows.

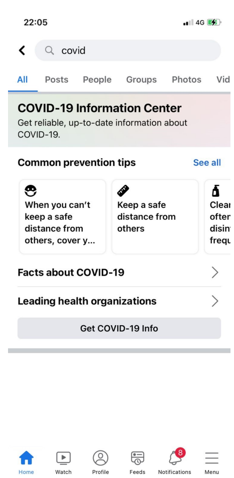


Figure 1.1: Social Media Warnings.

As it has been studied, fake news can diffuse farther and faster than truthful news, especially on political issues (5). Therefore, misinformation affects people during elections, with a percentage making civic decisions based on false knowledge (6; 7; 8). In general, unreliable information is present in our daily lives, and it becomes a challenge for technology to understand this phenomenon.

Contradictions represent one form of misleading information, for example, we can sometimes read a sensationalist title but the article can be unrelated to the title or even be contradictory to it. In other cases, politicians can incur contradictions when they are not informed on a subject. Or even, they can change their line of argument or opinion on opportunism.

Catching contradictions can be a first step in addressing this issue. So, processing the circulating information with machine learning can be a useful tool. Natural Language Processing (NLP) is the field of Machine Learning proper for this task since it aims to capture the meaning of a text in the natural language used by people. Some of the tasks that NLP solves are text classification, machine translation, and sentiment analysis, among others (9). Some of the applications for NLP tasks are spam filtering, question answering, and document summarization (10). One of the most notable achievements in NLP is chatGPT (11). This chatbot can answer questions that implicitly solve many tasks like summarizing web information of countless websites, solving mathematical problems, and creating raw code, among many others. It can also answer questions in many languages and answer follow-up questions inferring a context out of the previous queries.

This work implements a machine learning system that identifies contradictions. This system aims to classify the relationship between two sentences to detect whether the sentences are contradictory, compatible or there is no relation between them. This way, machine learning, and NLP techniques can become a useful tool to monitor and analyze circulating information.

2 | State of the art

2.1 Natural Language Processing

In Artificial intelligence, Natural Language Processing (NLP) is a field with many useful applications. Making a computer obtain information and insights from human language can be applied in different areas like language translation, social media monitoring, text mining, and customer care, among many others (12). Propelled by deep learning models, there are many NLP tasks of interest, including information extraction, text classification, question answering, and machine translation (13).

Deep Learning Algorithms work with numerical data as an input. Therefore, there is a need to convert natural and readable text to a numerical representation. One Hot Encoding is one option to represent words to be a valid input for machine learning models. This encoding makes vectors in N dimensions, where N is the number of words to consider. For each word, only one dimension has a unitary value, the rest are zeros. To represent every word a different dimension is activated (14).

Another option to approach this matter is “learning a distributed representation for each word, also called a word embedding” as Bengio states (15). The author argues that this kind of representation can discern underlying explanatory factors of the data. Furthermore, those previous insights captured by the embedding are useful when building a classifier or predictor (15). For example, a word embedding can represent features like semantic and syntactic information of the words which improves generalization for downstream tasks (16; 17).

One pre-trained word embedding is the Skip-gram model (18). This model encodes the linguistic patterns as simple algebraic operations. For example, the relationship “plural nouns” can be represented as C = vector(“mouse”) - vector (“mice”). The relationship C can be applied to other words and find the same relationship, for example, vector(“dollar”) + C allows us to find the vector for “dollars”. Many other relationships are encoded in this embedding like “adjective to verb”, “opposites”, “superlatives”, “past tense”, and “capital city” (18). Since this model is focused on words, it has limitations. For example, idiomatic phrases such as “New York Times” are not captured directly (19). Other embeddings consider whole sentences (20) or even paragraphs (21).

Concatenating embeddings that were trained on unlabeled data, and then fine-tuning on supervised data achieves good results so it is the preferred option nowadays. This methods were developed on different tasks like natural language inference, question answering, and machine translation, among others (22; 23; 24). In essence, these algorithms are used as a pre-trained model for then solving supervised learning problems (25).

BERT (Bidirectional Encoder Representations from Transformers) (22) is a language representation model that captures the context in both the left and right words of the sentence (bidirectional).

This model can be pre-trained on unlabeled data of the target language and then fine-tuned for different goals. The technique is described as adding a single layer for the specific NLP tasks and in this way it achieves a desired performance. An example of this is the application on sentiment analysis (26; 22) and contradiction detection (27).

2.2 Contradiction Detection

For the NLP task of contradiction detection, we must note that there are many contradiction types. Antonyms, Negation, Numeric, and Lexical, among others. Some types can be more complex to identify than others. For example, Structure and Lexical contradictions are the most difficult to detect as they require more lexical or world knowledge (28).

Specific word embeddings can be trained for a specific context. An example is CWE (Contradiction-specific Word Embedding) (29). This model allows contrasting words (antonyms) to be as distant as possible, and thus the model captures the semantic relation between words well in the context of contradiction detection. This technique has better performance on this task than previous traditional context-based word embeddings (29). In addition, research shows that introducing structured semantic information allows the model to improve performance on transfer learning techniques. An example is Semantics Aware BERT trained for Natural Language Inference (SNLI)(30).

For Spanish, the ES-Contradiction dataset was created with three categories: contradiction, entailment, and neutral (31). The technique previously mentioned is applied using a pre-trained model, in this case, BETO (BERT model adapted to Spanish) (32). This way the model is able to detect contradiction in four categories: negation, antonym, numeric, and structure. This model cannot detect accurately more complex categories as Lexical or World knowledge contradictions. Different approaches must be developed for the improvement in the detection of these kinds of contradictions.

Another task of research in this area is the detection of self-contradictory articles, specifically applied on Wikipedia articles. The main difference is that it takes the whole text as input to find a number of sentences that contradicts with each other. For this, a Pairwise Contradiction Neural Network (PCNN) is trained to achieve the state of the art results (33). This model is conformed by a stage that separates all sentences in an article and generates a representation for each using BERT. Then, in the second stage, all the representations are processed in pairs to detect contradictions. The final stage processes all the inputs to classify the whole text and predict if it is self-contradictory.

2.3 BERT

The first alternative to solve the problem of contradiction detection is the implementation of BERT fine-tuning for text classification. It has proven to be an efficient way to apply transfer learning from the already pre-trained BERT model to aim different NLP tasks (22). The proposed methodology in state-of-the-art work (26) focuses on sentiment analysis. That is, classify the sentence in positive, negative, or neutral emotion. For this, the BERT model was added a classifier layer and then trained on a dataset with the sentences and their respectively emotion labels (0,1,2).

To apply this technique to the specific problem of contradiction detection, two major changes must be done. The first change is to modify input structure. The proposed model currently accepts 1 sentence and must accept two sentences for our application. The second change consists on training the model (fine-tuning) with a different dataset. The original dataset was sentiment analysis oriented. The new dataset must be contradictions oriented.

Two datasets are proposed for this alternative. One is available on Kaggle.com (34). The second dataset is the XNLI cross-lingual corpus (27). Both datasets have for each pair of sentences the labels of contradiction, entailment, or neutral. Also, both datasets have samples of different languages that must be filtered to train in English only.

Two experiments are proposed. Experiment 1 consists in mixing both datasets to make a training and testing dataset. Experiment 2 consists on using only part of the Kaggle dataset to train. For testing a part of Kaggle and XNLI datasets are used (The same as Experiment 1).

The proposed methodology suggests a single layer for the classification task (26), but new architectures can be tested for the task. Finally, the main challenge this alternative presents is modifying the input structure and the effect it may present on the classifier results.

2.4 BETO

The second alternative focuses on fine-tuning the BETO model on the task of contradiction classification (31). In this case the task will be Spanish oriented. The methodology is similar to the previous alternative since BETO is based on BERT, but this model is pre-trained on a Spanish dataset generating a different representation space.

The main difference between both alternatives is the dataset. The dataset for this alternative will be composed by the ES-Headlines dataset (31) and the Spanish samples of the Kaggle dataset (34). Since the Kaggle dataset contains multilingual examples, a training dataset of only Spanish Kaggle samples is significantly smaller than the previous alternative. Therefore, the adding of ES-headlines dataset makes the training dataset comparable to previous alternative. Another option is to use the XNLI dataset (27). For this, each sample must be translated by a program or web service. This way, the quality of the dataset will depend on the quality of the translator service.

Two experiments are proposed for this alternative. Experiment 1 considers only the ES-Headlines dataset divided on training and testing datasets. Experiment 2 considers the spanish samples of Kaggle dataset combined with ES-Headlines dataset for training.

For this alternative, the model proposed in the previous work (31) is already structured to accept 2 sentences so no significant changes will be needed. Finally, similar to the previous alternative, the classifier layer can be restructured to try different approaches to the task.

2.5 Comparing alternatives

Both alternatives previously mentioned are compared on different points in order to decide which approach take for the task: language, dataset, changes to the model structure, and amount of related work on the specific language.

The first alternative offers the benefit that the amount of English datasets is vast compared to the amounts in Spanish. Also, we have experience with the model and code proposed. One challenge for this alternative is that the model structure must be modified to adapt to the task. For alternative 2, one benefit is that the model does not need much change as it is already structured on the task of contradiction detection. The challenges for this alternative are that there is less amounts of Spanish datasets in this task, so the application of this work on specific contexts requires the creation of a new dataset or translating the already existing in English.

	Alternative 1 BERT	Alternative 2 BETO
Language	English	Spanish
Model Structure	Requires adaptation	No major changes are required
Related Work	Abundant	Less abundant
Kaggle Dataset	~6000	~ 260

Table 2.1: Comparing alternatives

Table 2.1 summarizes the comparison between both alternatives. Alternative 2 is the one selected because usability is the most important factor that motivates this work. There is less work exploring the Spanish language compared to English, so it can be more useful to gain knowledge that can be applied in Latin America. Considering how innovative and useful the system and its applications can be, the amount of work published in the Spanish Language, and the possibility of testing on native speakers, BETO is the selected alternative for contradiction detection in Spanish.

3 | Methods

As previously mentioned, BETO was the selected alternative for solving the task of contradiction detection in Spanish. The main reason is that training a Spanish model can be more useful and convenient in our context which is Latin America. With more possibilities to apply this method for social or commercial purposes.

3.1 System

The proposed system is formed by two stages to solve the problem of contradiction detection. Stage A is responsible for speech recognition, that is transcribing audio into text. Stage B is responsible for contradiction detection, that is receiving the text and classifying the relationship between the sentences.

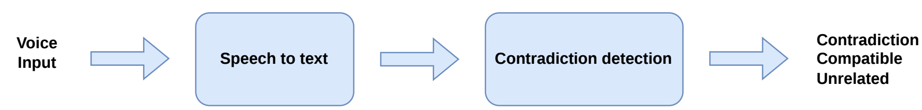


Figure 3.1: System design.

Stage A

This stage is in charge of Speech recognition, in other words, transforming the audio inputs of a microphone into readable text. All transcripts will be processed later in stage B.

Sentence input is manual. Audio input must be activated for each sentence by a key on the device. A customizable time of 15 seconds is considered to speak each sentence. After having entered both sentences, the system automatically goes to stage B.

To achieve this objective, the library SpeechRecognition for Python (35) is used. This library supports Google Speech Recognition and is used to transcribe audio. It also supports PyAudio (36) which takes the microphone input to the speech recognition function.

Stage B

This stage is in charge of classifying pairs of sentences into the three categories. The classes are contradiction, compatible, and neutral. To achieve this objective the model BETO will be fine-tuned to generate the multi-class classifier (31). BETO model is based on BERT and trained for natural language tasks in Spanish (32).

3.2 Architecture

The BERT model is based on transformers. The transformer is trained on the language translation task and its structure is purely based on attention mechanisms inside an encoder and a decoder (37). The encoder considers a stack of self-attention heads and a fully connected feed-forward network as figure 3.2 shows. The decoder considers a structure similar to the encoder. The decoder input is the encoder output. In addition, the decoder has a masked attention head for another input shown in figure 3.2. This last element allows the decoder outputs to re-enter but makes sure that only the previous outputs are considered to generate the next output. For instance, when translating the word in position n=3, only the encoder outputs that are in previous positions (n < 3) can re-enter as input to the encoder. This is shown in figure 3.3.

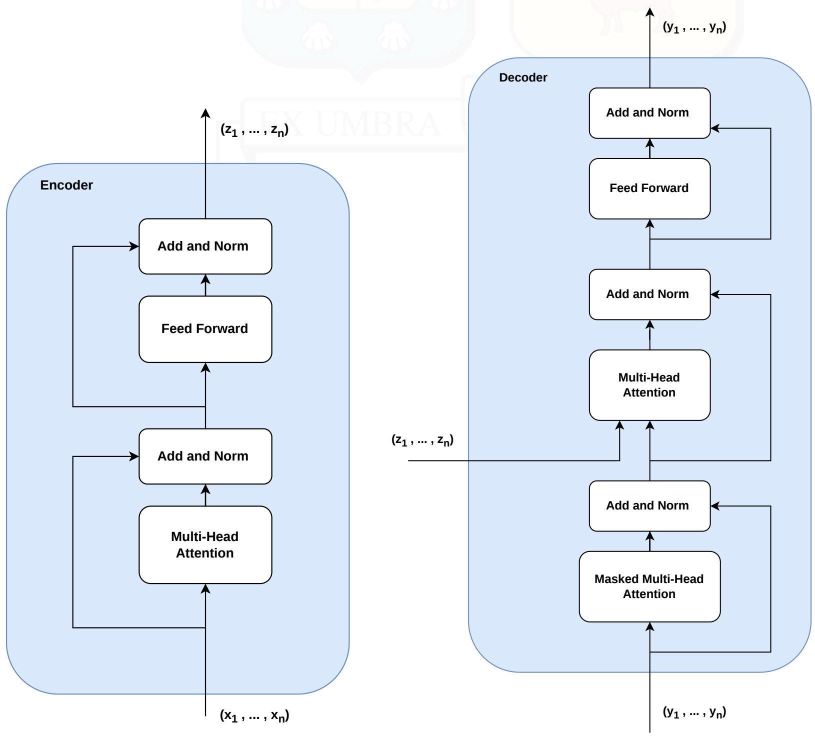


Figure 3.2: Transformer Architecture, Encoder and Decoder.

The model of this work is based on the code for BETO found in the following link (38), and uses libraries PyTorch (39) and Simple Transformers (40). As figure 3.4 shows, the model receives 2 tokens, CLS and SEP, along with the two input sentence tokens. CLS indicates the beginning of a sample and SEP is the separation between the two sentences. BERT was originally trained in masked language modeling and next-sentence prediction tasks and can be fine-tuned for other tasks, as figure 3.5 represents.

The BETO model was pre-trained in the Spanish corpus to be able to be used in this language. The classifier model is a fine-tuned version of BETO, trained on the contradiction detection task. It has 12 self-attention layers, 12 attention-heads on each layer, a hidden size of 768, and a sum of 110M parameters (31). The general methodology for this work is adapting the BETO model for classification. This adds a single linear layer on top of the pooled output of the original BETO model as figure 3.6 shows. This layer has as an output the class predicted. Only the classifier layer

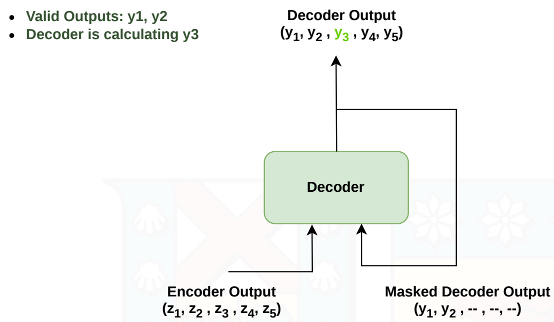


Figure 3.3: Transformer Masking.

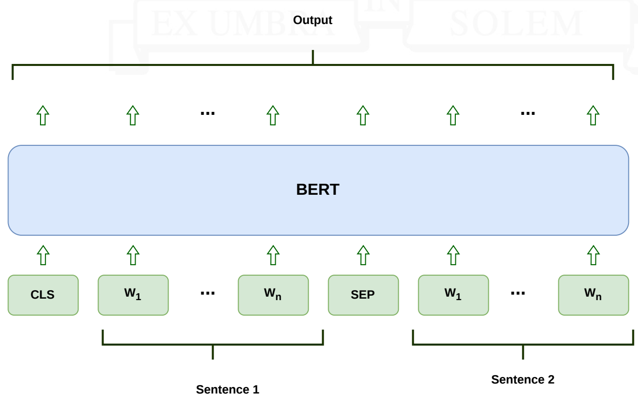


Figure 3.4: BERT architecture.

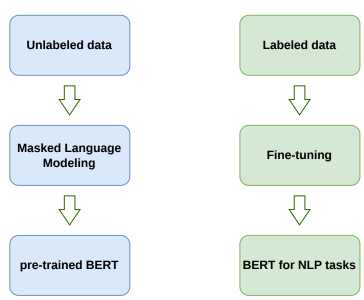


Figure 3.5: BERT operation.

is updated during training with this fine-tuning technique. This model is trained considering the following hyperparameters: learning rate of 2e-5, 3 epochs, and batch size of 4.

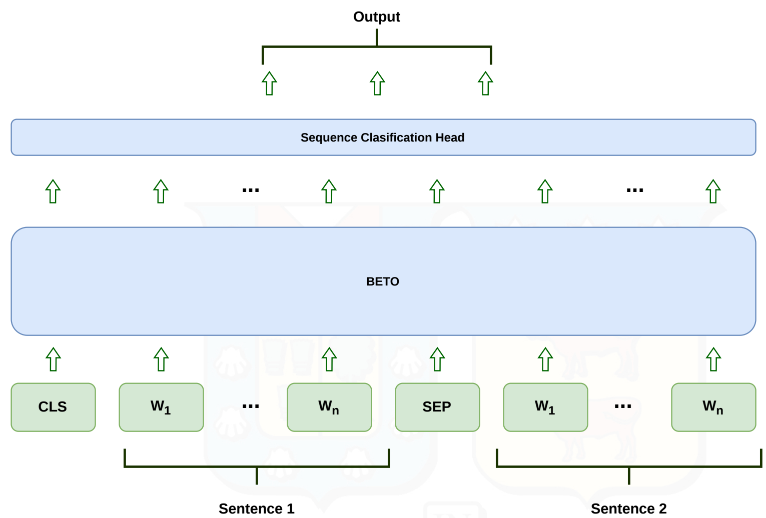


Figure 3.6: BETO for contradiction detection.

3.3 Datasets

Two datasets are used for training the model and only one is for testing and evaluating performance. The datasets are the Spanish Headlines Dataset (41) and the Kaggle dataset (34). The labels for each dataset are 0 for compatible texts, 1 for contradictory texts, and 2 for unrelated texts. It is important to note that the texts studied in the datasets have no slang or sarcasm. In addition, this work does not aim to detect what kind of contradiction is present, but only if it is a contradiction or not.

ES Headlines dataset

This is the original dataset from the studied paper (31). The objective of this dataset is to find contradictions between a news article and its headline, in the context of identifying news that might spread misinformation.

Each sample in the dataset consists of the headline, the corpus of the news article, and the label. The label has three possible values of interest: contradiction, compatible and unrelated. That value represents the relationship between the news article and its headline. The training dataset consists of 18168 samples and the testing dataset consists of 7777 samples.

Headlines+Kaggle dataset

The Kaggle dataset is the original dataset from the "Contradictory, My Dear Watson" Kaggle challenge (34). This dataset is focused on detecting contradictions in different contexts, not only news. It is also a multilanguage dataset so only the 259 Spanish samples were used. This dataset must be pre-processed to be a feasible entry into our system during training. A new training dataset will be formed (Headlines+Kaggle) that includes all samples from the ES-Headlines dataset and the 259 samples from the Kaggle dataset. The goal of this new dataset is to enrich the training of our model.

4 | Results

Two experiments were conducted during training to observe the effects of the dataset on the model's final performance. In both instances, only the classifier layer was updated during training.

Four metrics are employed for the analysis of the model performance. Precision and recall are computed for each class to evaluate model results. F1-score and accuracy are used to compare performance to the state-of-the-art.

The labels represented are: 0 - Compatible , 1 - Contradiction , 2 - Unrelated.

4.1 Experiment 1

This experiment considers training the model with the ES-Headlines Training dataset and Testing with the ES-Headlines Testing dataset. After training, the model accuracy is 0.93298. Table 4.1 presents the precision, recall, and F1 score for each class.

Label	Precision	Recall	F_1
0	0.90347	0.92258	0.91293
1	0.92108	0.90490	0.91292
2	0.99670	0.99179	0.99424

Table 4.1: Model performance Experiment 1.

4.2 Experiment 2

This experiment considers training the model with the Headlines+Kaggle Training dataset. The testing is conducted with the ES-Headlines Testing dataset like the previous experiment to evaluate and compare performance against the same parameters. The model accuracy is 0.93494 after training. Table 4.2 presents the precision, recall, and F1 score for each class.

Label	Precision	Recall	F_1
0	0.91095	0.92462	0.91774
1	0.92334	0.90987	0.91656
2	0.99288	0.99288	0.99288

Table 4.2: Model performance Experiment 2.

4.3 Comparison between state of the art and trained model

A comparison between state-of-the-art (31), experiment 1, and experiment 2 performance is presented below. Table 4.3 shows F1-score computed for each label and accuracy. Experiment 2 obtains a higher F1-score for the labels "Contradiction" and "Compatible". Experiment 2 also obtains a higher accuracy.

	F_1			Accuracy
	0	1	2	
Experiment 1	0.9129	0.9129	0.9942	0.9330
Experiment 2	0.9177	0.9166	0.9929	0.9349
State-of-the-Art	0.8870	0.8912	0.9959	0.9249

Table 4.3: Performance Comparison.

4.4 System operation

The following figures 4.1 and 4.2 show the designed system in operation. On the image appears the detected transcription of the audio for sentences 1 and 2 along with the system output (highlighted in purple). Of the six examples shown, three are contradictions, two examples correspond to unrelated sentences, and one is an example of compatible sentences.

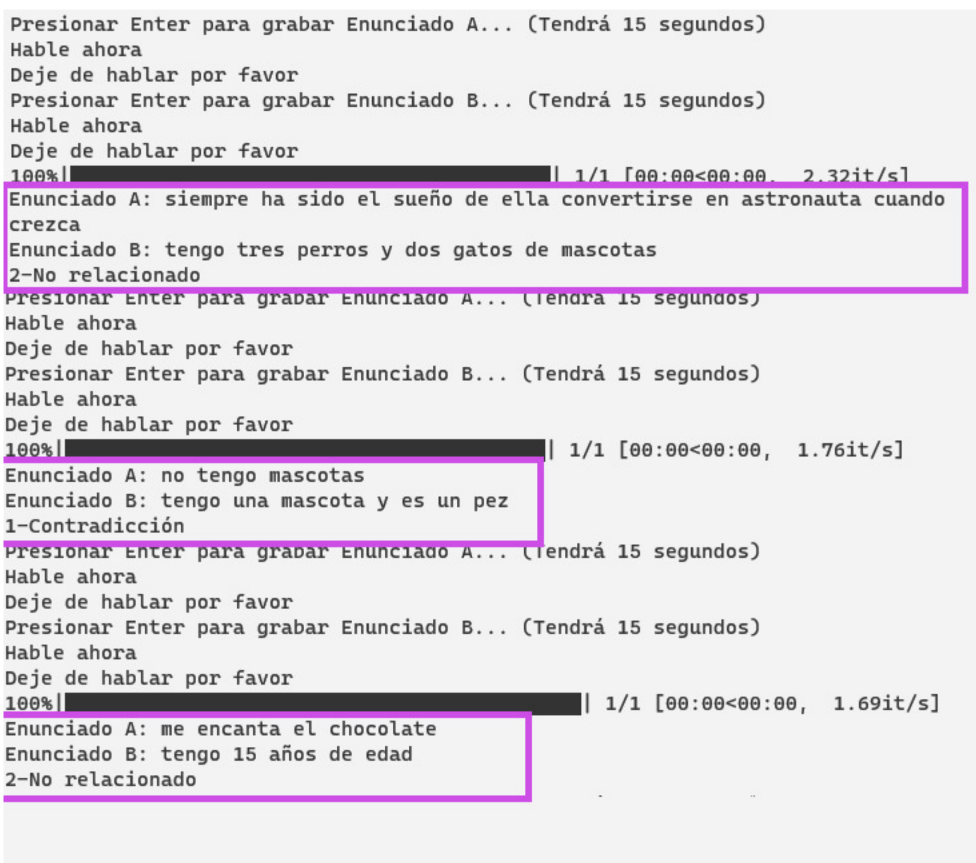
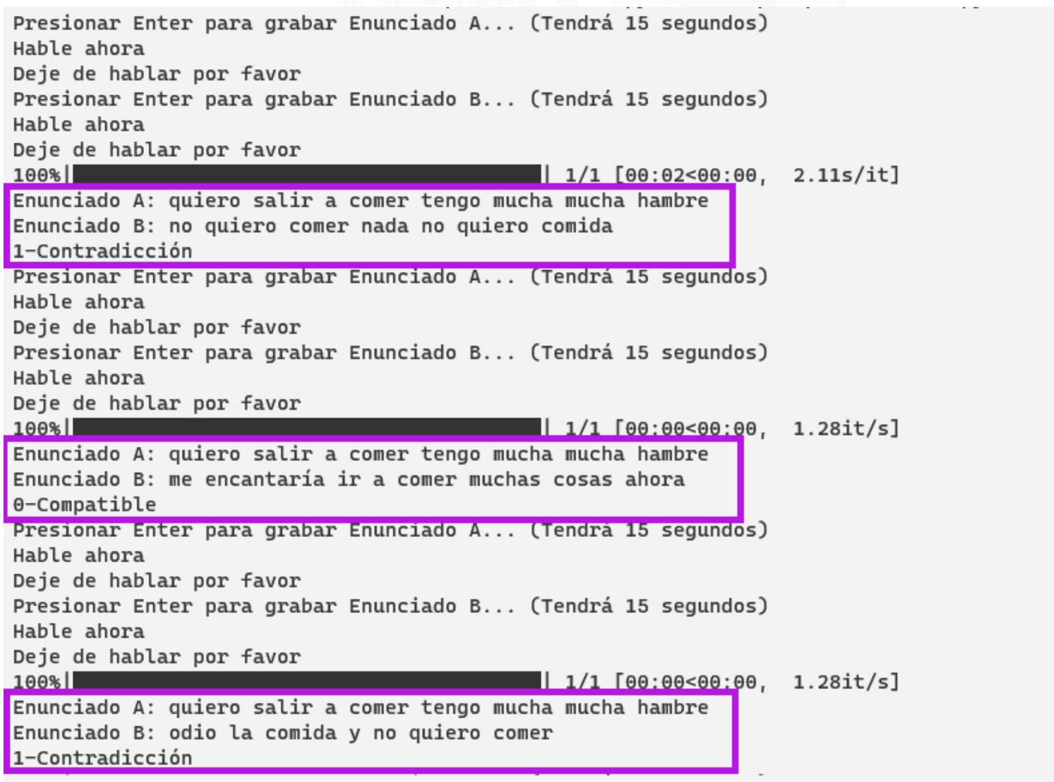


Figure 4.1: System running A.



```
Presionar Enter para grabar Enunciado A... (Tendrá 15 segundos)
Hable ahora
Deje de hablar por favor
Presionar Enter para grabar Enunciado B... (Tendrá 15 segundos)
Hable ahora
Deje de hablar por favor
100%| | 1/1 [00:02<00:00,  2.11s/it]
Enunciado A: quiero salir a comer tengo mucha mucha hambre
Enunciado B: no quiero comer nada no quiero comida
1-Contradicción
Presionar Enter para grabar Enunciado A... (Tendrá 15 segundos)
Hable ahora
Deje de hablar por favor
Presionar Enter para grabar Enunciado B... (Tendrá 15 segundos)
Hable ahora
Deje de hablar por favor
100%| | 1/1 [00:00<00:00,  1.28it/s]
Enunciado A: quiero salir a comer tengo mucha mucha hambre
Enunciado B: me encantaría ir a comer muchas cosas ahora
0-Compatible
Presionar Enter para grabar Enunciado A... (Tendrá 15 segundos)
Hable ahora
Deje de hablar por favor
Presionar Enter para grabar Enunciado B... (Tendrá 15 segundos)
Hable ahora
Deje de hablar por favor
100%| | 1/1 [00:00<00:00,  1.28it/s]
Enunciado A: quiero salir a comer tengo mucha mucha hambre
Enunciado B: odio la comida y no quiero comer
1-Contradicción
```

Figure 4.2: System running B.

5 | Conclusions

With this work, it is possible to solve the proposed problem of detecting contradictions in spoken language. For this a machine learning model based on BETO, the Spanish version of BERT was trained on labeled data.

Regarding the conducted experiments, it is observed that in the model resulting from the second training iteration (Experiment 2), the detection is better for more generic contradictions, including not only those associated with news. An accuracy increment of 0.2% is observed when increasing the training dataset in a 1.4% but focused on a different context.

It can be inferred that the performance can improve if the number of samples is increased. Furthermore, the same model can be re-trained to be applied to different situations and topics. A way to enrich the data for training for new contexts is data augmentation and the translation of existing English datasets to a language of interest, in this case, Spanish.

The results of this trained model are slightly higher than the original paper. It is inferred that this is due to the updates that have been made to the source code since the paper was published to when it was accessed to carry out this work.

Finally, with this work, we gained insights into how a model based on BETO or BERT identifies this type of relationship between texts. This kind of approach can be complemented with other works that identify fallacies or even determine if a statement is accompanied by a coherent argument or if it lacks support. All of this focused on the original objective of processing the information circulating on the internet and combating misinformation.

Bibliography

[1] M. McKee, M. C. I. van Schalkwyk, and D. Stuckler, “The second information revolution: digitalization brings opportunities and concerns for public health,” *European Journal of Public Health*, vol. 29, pp. 3–6, 11 2019. 1

[2] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018. 1

[3] “World Health Organization. Munich Security Conference.” https://www.who.int/director-general/speeches/detail/munich-security-conference, 2020. Accessed on August 29, 2023. 1

[4] “El coronavirus en la era de las redes sociales: De epidemia a ’infodemia’.” https://www.technologyreview.es/s/11887/el-coronavirus-en-la-era-de-las-redes-sociales-de-epidemia-infodemia, 2020. Accessed on August 29, 2023. 1

[5] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. 1

[6] E. Ferrara, “Disinformation and social bot operations in the run up to the 2017 french presidential election,” *First Monday*, jul 2017. 1

[7] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of Economic Perspectives*, vol. 31, pp. 211–36, May 2017. 1

[8] M. T. Bastos and D. Mercea, “The brexit botnet and user-generated hyperpartisan news,” *Social Science Computer Review*, vol. 37, no. 1, pp. 38–54, 2019. 1

[9] F. Chollet, “Deep learning with python,” 2017. P 310. 1

[10] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, pp. 3713–3744, January 2023. 1

[11] OpenAI, “ChatGPT.” https://chat.openai.com, 2023. Accessed on August 29, 2023. 1

[12] B. Priya, N. J.M, and G. Thangavel, *An Analysis of the Applications of Natural Language Processing in Various Sectors*. 10 2021. 2.1

[13] D. Otter, J. Medina, and J. Kalita, “A survey of the usages of deep learning for natural language processing,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, pp. 1–21, 04 2020. 2.1

[14] S. Bagui, D. Nandi, S. Bagui, and R. J. White, “Machine learning and deep learning for phishing email classification using one-hot encoding,” *Journal of Computer Science*, vol. 17, pp. 610–623, Jul 2021. 2.1

[15] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, pp. 1798–1828, 08 2013. 2.1

[16] J. Weston, F. Ratle, H. Mobahi, and R. Collobert, *Deep Learning via Semi-supervised Embedding*, pp. 639–655. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. 2.1

[17] F. Almeida and G. Xexéo, “Word embeddings: A survey,” 2023. 2.1

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013. 2.1

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” 2013. 2.1

[20] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, “Skip-thought vectors,” in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015. 2.1

[21] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.), vol. 32 of *Proceedings of Machine Learning Research*, (Bejing, China), pp. 1188–1196, PMLR, 22–24 Jun 2014. 2.1

[22] J. Devlin, M. Chang, K. Lee, and 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019. 2.1, 2.3

[23] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 670–680, Association for Computational Linguistics, Sept. 2017. 2.1

[24] B. McCann, J. Bradbury, C. Xiong, and R. Socher, “Learned in translation: Contextualized word vectors,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017. 2.1

[25] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015. 2.1

[26] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," in *China National Conference on Chinese Computational Linguistics*, 2019. 2.1, 2.3

[27] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: Evaluating Cross-lingual Sentence Representations," 2018. 2.1, 2.3, 2.4

[28] M.-C. de Marneffe, A. N. Rafferty, and C. D. Manning, "Finding contradictions in text," in *Proceedings of ACL-08: HLT*, (Columbus, Ohio), pp. 1039–1047, Association for Computational Linguistics, June 2008. 2.2

[29] L. Li, B. Qin, and T. Liu, "Contradiction detection with contradiction-specific word embedding," *Algorithms*, vol. 10, no. 2, 2017. 2.2

[30] Z. Zhang, Y. Wu, Z. Hai, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware bert for language understanding," in *AAAI Conference on Artificial Intelligence*, 2019. 2.2

[31] R. Sepúlveda-Torres, A. Bonet-Jover, and E. Saquete, ""here are the rules: Ignore all rules": Automatic contradiction detection in spanish," *Applied Sciences*, vol. 11, no. 7, 2021. 2.2, 2.4, 3.1, 3.2, 3.3, 4.3

[32] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish Pre-Trained BERT Model and Evaluation Data," in *PML4DC at ICLR 2020*, 2020. 2.2, 3.1

[33] C.-M. Hsu, C. te Li, D. Sáez-Trumper, and Y.-Z. Hsu, "Wikicontradiction: Detecting self-contradiction articles on wikipedia," *2021 IEEE International Conference on Big Data (Big Data)*, pp. 427–436, 2021. 2.2

[34] "Contradictory my dear watson." https://www.kaggle.com/competitions/contradictory-my-dear-watson, 2020. Accessed on August 29, 2023. 2.3, 2.4, 3.3

[35] A. Zhang, "Speech Recognition (Version 3.8)." https://github.com/Uberi/speech_recognition#readme, 2017. Accessed on August 29, 2023. 3.1

[36] H. Pham, "PyAudio." https://pypi.org/project/PyAudio/, 2006. Accessed on August 29, 2023. 3.1

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017. 3.2

[38] "ES-Contradiction baseline." https://github.com/rsepulveda911112/ES-Contradiction-baseline. Accessed on August 29, 2023. 3.2

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019. 3.2

[40] "SimpleTransformers Library." https://simpletransformers.ai/. Accessed on August 29, 2023. 3.2

[41] "ES-Headlines Dataset." https://zenodo.org/record/7602822. Dataset. Accessed on August 29, 2023. 3.3