applied sciences

MDPI

Article

"Here Are the Rules: Ignore All Rules": Automatic Contradiction Detection in Spanish

Robiert Sepúlveda-Torres *, Alba Bonet-Jover and Estela Saquete

Department of Software and Computing Systems, University of Alicante, Apdo. de Correos 99, E-03080 Alicante, Spain; alba.bonet@dlsi.ua.es (A.B.-J.); stela@dlsi.ua.es (E.S.)
* Correspondence: rsepulveda@dlsi.ua.es

Abstract: This paper tackles automatic detection of contradictions in Spanish within the news domain. Two pieces of information are classified as compatible, contradictory, or unrelated information. To deal with the task, the ES-Contradiction dataset was created. This dataset contains a balanced number of each of the three types of information. The novelty of the research is the fine-grained annotation of the different types of contradictions in the dataset. Presently, four different types of contradictions are covered in the contradiction examples: negation, antonyms, numerical, and structural. However, future work will extend the dataset with all possible types of contradictions. In order to validate the effectiveness of the dataset, a pretrained model is used (BETO), and after performing different experiments, the system is able to detect contradiction with a $F1m$ of 92.47%. Regarding the type of contradictions, the best results are obtained with negation contradiction ($F1m$ = 98%), whereas structural contradictions obtain the lowest results ($F1m$ = 69%) because of the smaller number of structural examples, due to the complexity of generating them. When dealing with a more generalistic dataset such as XNLI, our dataset fails to detect most of the contradictions properly, as the size of both datasets are very different and our dataset only covers four types of contradiction. However, using the classification of the contradictions leads us to conclude that there are highly complex contradictions that will need external knowledge in order to be properly detected and this will avoid the need for them to be previously exposed to the system.

Keywords: contradiction detection; natural language processing; deep learning; human language technologies

check for updates

Citation: Sepúlveda-Torres, R.; Bonet-Jover, A.; Saquete, E. "Here Are the Rules: Ignore All Rules": Automatic Contradiction Detection in Spanish. *Appl. Sci.* **2021**, *11*, 3060. https://doi.org/10.3390/app11073060

Academic Editor: Rafael Valencia-Garcia

Received: 6 March 2021
Accepted: 26 March 2021
Published: 30 March 2021

1. Introduction

One of the worst problems in the current information society is disinformation. It is a wide-ranging problem that alludes to the inaccuracy and lack of veracity of certain information that seeks to deliberately deceive or misdirect [1]. This phenomenon spreads on a viral scale and can therefore result in massive confusion about the real facts. Disinformation often involves a set of contradictory information that misleads users. Being able to automatically detect contradictory information becomes essential when the amount of information is so large that it becomes unmanageable and therefore confusing [2]. Contradiction, as described in [3], occurs between two sentences A and B when there exists no situation whatsoever in which A and B are both true. Therefore, in natural language processing (NLP), the task of contradiction identification implies detecting natural language statements conveying information about events or actions that cannot simultaneously hold [4]. In the current context, the automatic detection of contradictions would contribute to detect unreliable information, as finding contradictions between two pieces of information dealing with the same factual event would be a hint that at least one of the two pieces of news is false. A definition of different types of contradictions were presented in [3], where the authors defined a typology for English contradiction, finding two main categories: (1) those occurring via antonymy, negation, and date/number mismatch, which are relatively simple to detect, and (2) contradictions arising from the use of factive or modal words, structural and subtle lexical contrasts, as well as world knowledge (WK).

The task of automatic detection of contradictory information is tackled as a classification problem [5], when two pieces of text are talking about the same fact, within the same temporal frame. If we define a statement as $s = (i, f, t)$, where i refers to the information provided about fact f occurring at the time t, we will classify two pairs of text as

- *Compatible information*: two pieces of text, s_1 and s_2, are considered compatible if, given $s_1 = (i_1, f_1, t_1)$ and $s_2 = (i_2, f_2, t_2)$, the following statement holds true:

$$(i_1 \cong i_2) \wedge (f_1 = f_2) \wedge (t_1 = t_2) \tag{1}$$

- *Contradictory information*: two pieces of text, s_1 and s_2, are considered contradictory if, given $s_1 = (i_1, f_1, t_1)$ and $s_2 = (i_2, f_2, t_2)$, the following statement holds true:

$$(i_1 \ncong i_2) \wedge (f_1 = f_2) \wedge (t_1 = t_2) \tag{2}$$

- *Unrelated information*: two pieces of text, s_1 and s_2, are considered unrelated if, given $s_1 = (i_1, f_1, t_1)$ and $s_2 = (i_2, f_2, t_2)$, the following statement holds true:

$$f_1 \neq f_2 \tag{3}$$

Thus, a news item is classified as contradictory when given the same fact (It is considered that the same fact in two different news items could be expressed with different event mentions.) within the same time frame, the fact-related information is incongruent in the two news items being considered.

Nowadays, the coronavirus crisis has heightened both the need for reliable and not contradictory information. However, it is frequent to find different information about the same fact in different media, sometimes biased by a certain political spectrum. For example, here is a real case of contradiction in two different Spanish media outlets about the same information. The date of publication for the two news items taken from OkDiario and El Pais is the 19 March 2021:

1. Source "OkDiario" (https://okdiario.com/espana/estas-son-imagenes-reacciones-vacuna-astrazeneca-algunos-funcionarios-prisiones-6976588, accessed on 22 March 2021): *"Varios funcionarios han sufrido reacciones adversas tras la inoculación del fármaco que en algunos casos han precisado de atención hospitalaria por lo que piden un protocolo de seguimiento para los vacunados....Hasta ahora al menos tres policías nacionales, un guardia civil y un policía de la Ertzaintza han desarrollado trombos de gravedad tras haberse vacunado..."* ("Several government employees have suffered adverse reactions after being inoculated with the vaccine, and in some cases they have required hospital care, so they are calling for a follow-up protocol for those vaccinated...So far at least three national police officers, a civil guard and an Ertzaintza police officer have developed serious thromboses after having been vaccinated...")
2. Source "El Pais" (https://elpais.com/opinion/2021-03-19/confianza-en-las-vacunas.html, accessed on 22 March 2021): *"...La Agencia Europea del Medicamento ha ratificado que la vacuna de AstraZeneca es segura y eficaz y que los beneficios que aporta superan claramente a los posibles riesgos. Despeja así las dudas surgidas ante la notificación de una treintena de casos de trombosis..."* ("...The European Medicines Agency has confirmed that AstraZeneca's vaccine is safe and effective and that the benefits clearly outweigh the possible risks. This clears up the doubts that arose after the notification of some thirty cases of thrombosis...")

These two pieces of information concerning vaccination are contradictory, as while the first states that episodes of thrombosis have occurred after inoculation, the second rules out that a relationship exists between the cases of thrombosis that have occurred and the vaccine. This type of disinformation caused by the contradiction of information between the traditional media is potentially dangerous, as it may cause a public health problem generated by a reluctance to take up the offer of vaccination against COVID-19. Therefore, there is a need to alert users of these contradictions.

Most of the resources and systems for contradiction detection are developed in English [6–9]. However, despite the fact that Spanish is one of the most widely spoken languages in the world, there are no powerful resources to carry out the task of detecting contradictions from the direct perspective of this language. Currently, XNLI [10] is a cross-lingual dataset, which is divided into three partitions: training, developed, and test. The training set is developed in the English language, and the development and test sets are in 15 different languages. The XNLI has been used to create contradiction detection systems for training in English and predicting in other languages, obtaining good performance results. Each example in XNLI is classified as Contradiction, Entailment, or Neutral. However, to deal with contradictions it is important to consider their wide range and large variety of features [3]. Therefore, the purpose of this paper is to demonstrate that differentiating between the different types of contradictions can help to perform a more specific treatment of them, thereby enhancing capability to detect them in a broader way without having many previous examples of them. The XNLI dataset does not distinguish between different types of contradictions in its annotation, and the Spanish language is only available in the development and test sets manually translated from English. Both of these facts may affect the performance of models created from XNLI dataset for different languages. Besides, in this sense, the novelty of the proposed work is that we focus the proposal beyond covering the detection of the contradiction in Spanish, towards being able to detect what type of contradiction it is.

Furthermore, the contradiction detection system can be applied to detect different types of disinformation such as incongruent headlines or news published by different media, whether traditional or social, that seek to inform about the same fact but the information provided is inconsistent, and thereby inaccurate and unreliable.

The main contributions of this research are the following:

- First, as there is a lack of Spanish resources created from scratch for this task, a new Spanish dataset is built with different types of compatible, contradictory, and unrelated information for the purpose of creating a language model that is capable of automatically detecting contradictions between two pieces of information in this language. The novelty of this dataset and what differentiates it from others is the fact that in addition to detecting contradictions, each contradiction is annotated with a fine-grained annotation, differentiating between different types. Specifically, four of the types of contradictions defined in [3] are covered: antonymy, negation, date/number mismatch, and structural. In addition, the dataset is based on the study of incongruent headlines in traditional media, and it contains different types of contradictions between headlines and body texts in the Spanish language.
- Second, a set of experiments using a pretrained model as BETO [11] has been applied to build the language model and validate its effectiveness.

Note that at this stage of the research, covering only four types of contradictions is a real limitation of our dataset due to the wide spectrum of contradictions existing between texts. However, it allows the structure and design of preliminary systems for detecting contradiction in Spanish. The creation of an automatic process for classifying contradictions between texts, scaling from trivial to complex cases, could contribute to the design of hybrid systems operating in human–machine environments, providing additional information to humans about the type of contradiction encountered in an automatic system, which is the future line of our research.

The rest of the paper is organized as follows. Section 2 describes the previous work and existing resources on contradiction. Section 3 presents the definition of the dataset benchmark. Section 4 describes the model, the evaluation setup used, and experiments conducted in this research. Section 5 presents the results and discussion. Finally, our conclusions and future work are presented in Section 6.

2. Related Work

In this section, a brief review of contradiction detection methods is presented. Besides, some research in the field for specific domains is introduced. Finally, because the most important aim of this research is the creation of a new dataset for this field, a review of the main existing resources is provided below.

2.1. Contradiction Detection Methods

- Linguistic features approaches

The most common approaches to contradiction detection in texts use the linguistic features extracted from texts to build a classifier by training from the annotated examples, such as the works in [3,5,12,13].

Early research on contradiction detection within the field of natural language processing was reported by the authors of [12] whose work tackled contradictions by means of three types of linguistic information: negation, antonymy, and semantic and pragmatic information associated with discourse relations. After evaluation experiments, over 62% of accuracy was obtained due to the fact that there are more types of contradictions possible in texts.

Linguistic evidences such as polarity, numbers, dates and times, antonymy, structure, factivity, and modality features were used by the authors of [3] to detect contradiction.

An approach for detecting three different types of contradiction (negation, antonyms, and numeric mismatch) was proposed in [5]. This approach deploys a Recurrent Neural Network (RNN) using long short-term memory (LSTM) and Global Vectors for Word Representation (GloVe) and included four linguistic features extracted from the text: (1) Jaccard Coefficient, (2) Negation, (3) IsAntonym, and (4) Overlap Coefficient.

Simple text similarity metrics (cosine similarity, f1 score, and local alignment) were used as baseline in [13], obtaining good results for contradiction classification. This approach used two datasets built with examples of tweet pairs.

- Semantic features approaches

Other approaches created contradiction detection systems are based on semantic features [4,14–16].

A model to detect contradiction and the architecture that enables validation of the model was proposed by [4]. The model defined the extraction of semantic relations between a pair of sentences and verified some rules to detect contradictions. Furthermore, this author defined contradiction measures by considering the structure of relations extracted from texts and the level of uncertainty attached to them.

Other authors [14] combined shallow semantic representations derived from semantic role labeling (SRL) with binary relations extracted from sentences in a rule-based framework, and the authors of [15] extended the analysis using background knowledge.

A contradiction-specific word embedding (CWE) model and a large-scale corpus of contrasting pairs were proposed in [16]. This approach improved the results in contradiction detection in SemEval 2014 [17]. This research concluded that traditional word embedding learning algorithms have been highly successful in accomplishing the main NLP tasks but most of these algorithms are not powerful enough for the contradiction detection task [16].

2.2. Contradiction Detection in Specific Domains

There is also some specific domain research regarding contradiction detection. In medical domain, the authors of [18] detected contradiction by comparing subject–relation–object tuples of a text pair in medical research. This work detected 2236 contradictions automatically, but these contradictions were checked manually and only 56 were correct.

A classification system based on Support Vector Machine (SVM), with some features (negation, antonyms, and similarity measures) that help to detect contradiction in medical texts was created in [19]. This system detected antonyms and negation contradiction but not numerical contradiction. These results improved the state-of-the-art in a medical dataset.

Regarding the tourism domain, other research provides an analysis of the type of contradictions present in online hotel reviews. In addition, a model for the detection of numerical contradiction is proposed for the tourism industry [20].

2.3. Contradiction Detection Resources

Currently, the availability of large annotated datasets for contradiction detection are mainly present in English [21], such as SNLI [6], MultiNLI (including multiple genres) [7], or even the cross-lingual dataset XNLI [10]. These datasets have allowed the training of complex deep learning systems, which require very large corpora to obtain successful results. There are numerous studies that use these resources to create Recognizing Textual Entailment (RTE) systems. These systems usually use Transformers Learning models like BERT [22] and RoBERTa [23] to improve their predictions. BERT and RoBERTa are multi-layer bidirectional Transformer encoders that are designed to pre-train from text without labels. These pretrained models have the advantage of being able to be fine-tuned with just one additional layer of output, a feature that enables them to be used to create state-of-the-art models in various NLP tasks.

In addition, there is research that merges deep learning models with external knowledge. The Wordnet relations were introduced in [8] to enrich neural network approaches in natural language inference (NLI), which is a previous step in contradiction detection. In another sense, the research developed in [9] introduces SRL information that allows the improvement of models based on Transfer Learning.

To the authors' knowledge, there are few studies that address the detection of contradictions in languages other than English, such as those in [21,24]. Machine translation of SNLI from English into German was done in [21]. They built a model on the German version of SNLI and the results of the predictions are very similar to the same model trained on the original SNLI version in English. A large-scale database of contradictory event pairs in the Japanese language has been created by [24]. This database is used to generate coherent statements for a dialogue system.

As for multilinguality, current research in NLI is mainly conducted in English. Concerning other languages, cross-lingual datasets were provided in [25] and XNLI in [10]; however, they relied on translation-based approaches or multilingual sentence encoders. The detection of contradictions is a very complicated task within the NLP [21]. It would be convenient to have powerful datasets in Spanish that allow the creation of specific systems to detect contradictions in Spanish. Furthermore, existing datasets do not determine the different types of contradictions, whereas considering a fine-grained annotation in the contradictions would be more effective for dealing with them. Given these considerations, one of the main aims of this work is the development of a Spanish dataset that contains a balanced number of compatible, contradictory, and unrelated information in a first step, and subsequently, differentiating the different types of possible contradictions. The process followed to build the dataset is described in detail in the next section.

3. ES-Contradiction: A New Spanish Contradiction Dataset

Our dataset (ES-Contradiction) is focused on contradictions that are likely to appear in traditional news items written in the Spanish language. Unlike other datasets, in the dataset proposed in this work, contradictions are annotated by distinguishing the type of contradiction according to its specific characteristics. Thanks to this fine-grained classification, complex contradictions can be treated in more precisely in future.

In order to create the ES-Contradiction dataset, news articles from a renowned Spanish source were automatically collected, including the headline and body text. According to the journalistic structure of a news item, the headline is the title of the news article, and it provides the main idea of the story. Normally, in one sentence it summarizes the basic and essential information about the story. The main objective of the title is to attract the reader's attention. A headline is therefore expected to be as effective as possible, without losing accuracy or becoming misleading [26]. Therefore, finding contradictions between headlines and body texts is a crucial task in the fight against the spread of disinformation.

In the current state of the dataset, news is focused on two domains—economics and politics, although the ultimate goal will be automatic cross-domain contradiction detection.

3.1. Dataset Annotation Stages

The dataset was built in four stages, subsequently outlined and detailed: (1) Extracting information from data source, (2) modifying news headline according to the different types of contradictions, (3) classifying the relationship between headline and body text (Compatible or Contradiction), and (4) randomly mixing headlines and body texts (Unrelated).

1. Extracting information from data source: The headline, body text, and date of the news item are extracted from a reliable data source. In this case, the news agency EFE was used (https://www.efe.com/efe/espana/1, accessed on 22 March 2021). The news extracted belongs to the political and economic domains, assuming that the headlines and body texts are compatible, although in the third stage this relationship is verified.
2. Modifying news headlines: The aim of this stage is to make the news headline contradictory to the body text by including simple alterations to the headline structure. The changes to the headline together with some examples are (examples given in Spanish and translated into English for clarification) given as follows:
   - NEGATION (Con_Neg): This alteration consists of negating the headline of the news item.
     (a) Original headline: *"El comité de empresa* debatirá *mañana la "propuesta final" de Alcoa"* ("Union representatives *will discuss* Alcoa's 'final proposal' tomorrow")
     (b) Modified headline: *"El comité de empresa* no debatirá *mañana la "propuesta final" de Alcoa"* ("Union representatives *will not discuss* Alcoa's 'final proposal' tomorrow")
   - ANTONYM (Con_Ant): This transformation consists of replacing the verb denoting the main event of the headline with an antonym.
     (a) Original headline: *"El Gobierno se compromete a* subir *los salarios a los empleados públicos tras los comicios"* ("The Government pledges to *raise* public employees' salaries after the elections")
     (b) Modified headline: *"El Gobierno se compromete a* bajar *los salarios a los empleados públicos tras los comicios"* ("Government pledges to *cut* public employees' salaries after the elections")
   - NUMERIC (Con_Num): This amendment consists of changing numbers, dates, or times appearing in the headline.
     (a) Original headline: *"La economía británica ha crecido un* 3% *menos por el brexit, según S&P"* ("UK economy has grown by *3%* less due to Brexit, says S&P")
     (b) Modified headline: *"La economía británica ha crecido un* 5% *menos por el brexit, según S&P"* ("UK economy has grown by *5%* less due to Brexit, says S&P")
   - STRUCTURE (Con_Str): This modification consists of changing the position of one word for another or substituting words in the sentence.
     (a) Original headline: "Arvind Krishna *sustituirá a* Ginni Rometty *como consejero delegado de IBM"* ("*Arvind Krishna* will replace *Ginni Rometty* as IBM's CEO")

(b) Modified headline: "Ginni Rometty *sustituirá a* Arvind Krishna *como consejero delegado de IBM*" ("*Ginni Rometty* will replace *Arvind Krishna* as IBM's CEO")

These alterations will change the semantic content of the sentence, making it contradictory to the previous headline and body text. The annotation process was carried out by two independent annotators that were trained by an expert annotator.

3. Classifying the relationship between the headline and the body text: The semantic relationship between the headline and the body text was annotated in two phases: The first phase consisted of classifying the information into Compatible (compatible information) or Contradiction (contradictory information). In the second phase, in the case of Contradiction, the type of contradiction was also annotated (Negation, Antonym, Numeric, Structure). This stage involved four annotators who are trained to detect semantic relationships between pairs of texts.
4. Aleatory mixing headline and body text: The news items reserved in the first stage were used to generate unrelated examples (Unrelated). The headline was separated from the corresponding body text and all the headlines were randomly mixed with the body texts. In the mixing process, it was verified that the headline is not mixed with the corresponding body text. This step was done automatically without the intervention of the annotators.

3.2. Dataset Description

The dataset consists of 7403 news items, of which 2431 contain Compatible headline–body news items, 2473 contain Contradictory headline–body news items, and 2499 are Unrelated headline–body news items. This represents a balanced dataset with three main classification items. The dataset split sizes for each annotated class are presented in Table 1. We partitioned the annotated news items into training and test sets.

Table 1. Dataset split sizes for each class.

Split	Compatible	Contradiction	Unrelated
Training	1703	1733	1755
Test	728	740	744
Total items	2431	2473	2499

As can be seen in Table 2, our dataset contains examples of each type of contradiction. However, it is important to clarify that there are few examples of structure contradiction, given the complexity of finding sentences that allow for this type of modification.

Table 2. Contradiction types in the dataset.

Split	Con-Neg	Con-Ant	Con-Num	Con-Str
Training	674	552	430	77
Test	287	236	184	33
Total items	961	788	614	110

3.3. Dataset Validation

Due to the particularities of the dataset annotation process, it was necessary to validate the second and third stages of the process. For the second stage, a super-annotator validation was conducted, while for the third stage, an inter-annotator agreement was carried out. We randomly selected 4% of the Compatible and Contradiction pairs (n = 200) to carry out the dataset validations.

3.3.1. Super-Annotator Validation

For the second stage, it was not possible to make an inter-annotator agreement because this stage consists of headline modifications and the possible variations are infinite. In this case, a manual review of the modified headlines is performed by the Super-Annotator to detect inconsistencies with the indications in the annotation guide. Only 2% of the analyzed examples present inconsistencies with the annotation guide, corroborating the validity of this stage.

3.3.2. Inter-Annotator Agreement

In order to measure the quality of the third stage annotation, an inter-annotator agreement between two annotators was performed. In cases where there was no agreement, a consensus process was carried out among the annotators. Using Cohen's $kappa$ [27] a $k = 0.83$ was obtained, which validates the third-stage labeling.

4. Experiments and Evaluation Metrics

A system capable of detecting contradictions is highly relevant as it would enable the improvement and support of other tasks that involve detecting contradictory pairs (fact-checking or stance detection). To test the validity of the newly created Spanish contradiction dataset in this task, a baseline was created that is based on the BETO (https://github.com/dccuchile/beto, accessed on 22 March 2021) model described in [11] that was previously pretrained in a Spanish dataset. Wikipedia texts and all OPUS Project sources [28] with Spanish texts are used as training data. The model used is based on the BERT [22] model, and it performs a series of optimizations similar to those performed in the RoBERTa model [23]. As with the BERT model, the input sequence to the model is the headline text concatenated with the body text.

The flexibility provided by BERT-based models allows us to create competitive baselines by fine-tuning the model on the dataset to be predicted [22].

4.1. Experimental Setup

The model was implemented using the Simple Transformer (https://simpletransformers.ai/ (accessed on 3 March 2021)) and PyTorch (https://pytorch.org/ (accessed on 3 March 2021)) libraries. In our experiments, the hyperparameter values of the model are maximum sequence length of 512, batch size of 4, training rate of 2e-5, and training performed for 3 epochs. These values were established after the cross-validation experiment (see Section 5.2).

4.2. Experiments

The main objective of the experimentation proposed in this research is to demonstrate that a model is able to learn how to automatically detect contradiction types and contradictions with high accuracy from the ES-Contradiction dataset.

The BETO model has been configured as indicated in Section 4.1, and the following experiments were performed:

1. Predicting all classes: This experiment uses ES-Contradiction dataset and the model is trained with the training set, resulting in the prediction of the test set. The classes to predict are Compatible, Contradiction, and Unrelated.
2. K-fold cross-validation: This experiment makes a cross-validation with our training set without unrelated examples. Cross-validation is a statistical technique that involves partitioning the data into subsets, training the data on a subset, and using the other subset to evaluate the model's performance. Cross-validation enables all available data to be used for training and testing [29]. In this experiment, k-fold cross-validation with k = 5 is used.
3. Detecting contradiction vs. compatible information: This experiment focuses on detecting only the contradictory and compatible examples from our dataset (Table 1 without the unrelated examples). The classes to predict are Compatible and Contradiction.

4. Detecting specific type of contradictions: This experiment uses only the contradictory examples of the dataset described in Table 2 to detect the types of Contradiction between pairs. The training and test set are used for training and testing.
5. Comparison between XNLI and our dataset: This experiment trains by using machine translation into Spanish of the XNLI dataset (https://github.com/facebookresearch/XNLI, accessed on 29 March 2021), and uses the Spanish test set of the XNLI corpus and our test set. The XNLI dataset has 3 classes: (Entailment, Contradiction, and Neutral). Therefore, it was necessary to match them with our dataset. The Neutral class of the XNLI dataset and the Unrelated class of our dataset were eliminated, whereas the Entailment class was associated with our Compatible class and the Contradiction class with our Contradiction class.

4.3. Evaluation Metrics

In order to evaluate the experiments, both a measure of F_1 class-wise and a macro-averaged F_1 (F_1m) as the mean of those per-class F scores are used, which also enables the imbalance among the less represented classes to be addressed. The advantage of this measure is that it is not affected by the size of the majority class. Additionally, accuracy (Acc) is also obtained.

5. Results and Discussion

This section presents the results obtained in each of the experiments described in Section 4. The values are expressed in percentage mode (%).

5.1. Predicting All Classes

This experiment is performed on the entire dataset to predict the 3 classes previously defined. The system created is capable of detecting the Unrelated class with a high level of precision and achieves significantly good results in the Compatible and Contradiction classes. Table 3 presents the results.

Table 3. Results obtained from Experiment 1: Predicting compatible, contradictory, and unrelated information.

		F_1 Score (%)		F_1m (%)	Acc (%)
System	**Compatible**	**Contradiction**	**Unrelated**		
BETO-All classes	88.70	89.12	99.59	92.47	92.49

The results obtained in the Unrelated class indicate that the system is capable of detecting with excellent F_1m these types of examples, corroborating the results obtained in the literature on this type of semantic relation between texts [30]. The other two classes have room for improvement, by using, for instance, external knowledge. A future line of work would consist of including resources that detect antonyms and synonyms in line with [31] for the purpose of improving the results of the Contradiction class. Furthermore, including syntactic and semantic information could improve the detection of other more complex contradictions, such as structural ones, without the need for such large datasets.

5.2. K-Fold Cross-Validation

A k-fold cross-validation experiment aims to estimate the error and select the hyper-parameters of the model [29]. This is achieved by training and testing the model with all available data for training. Table 4 shows the results of the cross-validation for each fold.

The experiment conducted with our best fine-tuning model obtains a mean accuracy of 88.94% and a standard deviation of 1.234%. The prediction of the contradiction classification model in the test set should have an accuracy close to the mean obtained in the cross-validation because the standard deviation is very low. Furthermore, the training and test set of the ES-Contradiction are very similar as they were formed by splitting the original dataset.

Table 4. Results obtained from Experiment 2: k-fold cross-validation.

	F_1 Score (%)		F_1m (%)	Acc (%)
Fold	**Compatible**	**Contradiction**		
1	90.03	90.75	90.39	90.40
2	90.43	89.43	89.93	89.95
3	88.69	88.88	88.79	88.79
4	86.64	88.21	87.43	87.48
5	87.76	88.35	88.05	88.06

5.3. Detecting Contradiction vs. Compatible Information

In this experiment, the Unrelated class is removed from the ES-Contradiction dataset to measure the accuracy of the approach in terms of distinguishing between compatible or contradictory information, assuming that the information is related. The results are shown in Table 5. The approach obtains similar results in both predicted classes. This is due to the quality of the training examples and the balanced number of examples from each class in this dataset. As indicated in the discussion of the first experiment, the results for predicting classes could be improved by introducing external semantic information, similar to the introduction of SRL [9] and the use of Wordnet relations [8], both of which improve the results of deep learning models.

Table 5. Results obtained from Experiment 3: Detecting between compatible and contradictory information when the texts are related.

	F_1 Score (%)		F_1m (%)	Acc (%)
System	**Compatible**	**Contradiction**		
BETO-Contra_Comp	88.63	88.75	88.69	88.69

5.4. Detecting Specific Types of Contradictions

This experiment aims to analyze the detection capability of the approach by contradiction types. Table 6 shows the results obtained exclusively for the detection of contradiction types.

Table 6. Results obtained from Experiment 4: Detecting each specific type of contradiction treated.

	F_1 Score (%)				F_1m (%)	Acc (%)
System	**Con-Neg**	**Con-Ant**	**Con-Num**	**Con-Str**		
BETO-Type of contradictions	97.90	93.20	92.39	68.75	88.06	93.78

The structural contradiction class (Con_Str) is the one that obtains the lowest accuracy results and F_1m. This contradiction type is considered one of the most complicated contradictions to detect compared with the other contradictions [3], which is in line with our results. In addition, the Con_Str class, due to the scarcity of training examples, contains the lowest number of examples in this dataset, so the model can learn more about other more representative classes. It is highly likely that contradictions such as the structure contradiction need external semantic knowledge to improve detection results.

5.5. Comparison between XNLI and ES-Contradiction

In order to demonstrate the generality of our proposal, a series of experiments has been performed using the XNLI dataset and ES-Contradiction dataset in different training–test configurations.

The XNLI dataset is divided into training, development, and test set. The training set is developed in the English language. The development and test sets are in 15 languages, including Spanish.

To carry out this experiment, machine translation of the training set into Spanish and the test set in Spanish were used. Table 7 presents the results of each trained system. The best results are highlighted in italic.

Table 7. Results obtained from Experiment 5: Results of different training-test configurations between XNLI and ES-Contradiction.

	F_1 Score (%)		**F_1m (%)**	**Acc (%)**
System	**Compatible**	**Contradiction**		
Training set (XNLI) and test set (ES-Contradiction)	72.57	75.00	73.78	73.84
Training and test set (XNLI)	71.73	77.53	74.63	74.97
Training set (ES-Contradiction) and test set (XNLI)	65.21	32.28	48.75	54.04
BETO-Contra_Comp	88.63	88.75	88.69	88.69

The models of line 1 and 2 are trained using the XNLI training set, the difference being that the first line predicts the ES-Contradiction dataset test set, and the other one, the XNLI test set. The prediction results are quite close for both of them, but the Contradiction class is detected with a higher accuracy and F_1m.

Comparing lines 1 and 4, considering our dataset as the test set with the four types of contradictions, as expected the system trained on our dataset is substantially better than the system trained on the XNLI training set. The result indicates that the XNLI dataset does not manage to cover all the contradictions contained in our dataset, even though it is more than 40 times the size of the training set of ES-Contradiction dataset and is composed of examples from different genres.

The XNLI training set is exactly the same as the MultiNLI training set. It has been developed manually by parsing a sentence from a non-fiction article and creating three sentence variants: definitely correct, might be correct, and definitely incorrect [7]. The procedure for creating the training set of MultiNLI dataset follows an annotation guide that is sufficiently general to avoid bias in the dataset. However, this lack of specificity may cause a shortage of examples of various types of contradiction, resulting in an imbalance of contradiction types. Table 8 shows the accuracy by type of contradiction of the model trained in row 1 of Table 7.

Table 8. Accuracy obtained for detecting each specific type of contradiction with the model in row 1 of Table 7.

	Acc (%)			
System	**Con-Neg**	**Con-Ant**	**Con-Num**	**Con-Str**
Training set (XNLI) and test set (ES-Contradiction)	88.50	75	70.10	48.48

In the prediction of the type of contradiction (Con_Neg, Con_Ant, and Con_Num), this model achieves significantly good results; even in the class Con_Neg they are very good (88.50% accuracy). However, in the prediction of the class Con_Str, they are very low (48.48% accuracy); this result could be due to the lack of examples of this type in the XNLI dataset.

Finally, the system trained on the ES-Contradiction dataset failed to obtain good enough results in order to predict the XNLI test set. This system only obtains 32.28% F_1m to predict the contradictions of the XNLI test set. The need to include new types of contradictions in the ES-Contradiction dataset is evidenced, specifically those that allow the creation of robust contradiction detection systems for the real-world and enable prediction with higher accuracy in the XNLI dataset.

Unlike the XNLI dataset, the ES-Contradiction dataset in its first version could not be used to create a real system of contradiction detection. However, the annotation of contradiction types has enabled us to detect which contradictions are more difficult to tackle and how models may need external knowledge to improve the results. By future inclusion of other types of contradiction in our dataset (factive, lexical, WK, and more examples of structure contradiction), we could assess what kind of knowledge is useful to include in the reference models within this task, and thereby make progress towards the creation of a powerful system for detecting contradictions.

Extending the XNLI dataset with the types of contradictions contained in the ES-Contradiction dataset is not an appropriate option as the XNLI Spanish language training set is automatically translated, which could incorporate several biases into automatic detection systems. Furthermore, the currently annotated examples do not have this fine-grained annotation of our proposal.

6. Conclusions

This work has built the ES-Contradiction dataset, a new Spanish language dataset that contains contradiction, compatible, and unrelated information. Unlike other datasets, in the ES-Contradiction dataset, contradictions are annotated with a fine-grained annotation that distinguishes the type of contradiction according to its specific characteristics. The contradictions currently covered in the dataset created are negations, antonyms, date/numerical mismatch, and structural contradictions. However, all the contradictions presented in [3] are the final goal of this research. The main purpose is to create an automatic process for classifying contradictions between texts, scaling from trivial to complex cases, and giving each contradiction a precise and customized treatment. This would avoid the need to have large datasets that contemplate a multitude of examples for each of the contradictions.

BETO model is used to create our system. Beto is a Transfer Learning model based on BERT. Five different experiments were performed with our system indicating that it is able to detect the four types of contradictions with a F_1m of 92.47% and contradiction types with a F_1m of 88.06%. As for the detection of each specific type of contradiction, our system obtains the best results for negation contradictions (98% F_1m), whereas the lower results are obtained for structural contradictions (69% F_1m), corroborating that the best results are obtained from the classes with the largest number of examples with more simple contradictions. Our results leave a great margin for improvement that can be tackled with the inclusion of external knowledge that enables improvement on the prediction of contradiction types.

Furthermore, as for the generalization of the system, we compared the system by training it on the XNLI dataset and training it on ES-Contradiction dataset. The system trained on our dataset was not able to detect with high accuracy the XNLI test set, which indicates that in this first version it is not possible to create a powerful contradiction detection system. The negative results in the generalization tests of our corpus were expected, as it only covers four types of contradictions existing in texts. On the other hand, the system trained on the XNLI dataset managed to detect the contradictions in our dataset with high accuracy, especially in the most common types of contradictions, which therefore will also be the largest number of examples. However, when analyzing by contradiction types, we detected that the structure contradiction is not detected correctly. With this experiment, we found that the XNLI dataset, although much larger than ours, does not cover all types of contradictions, which indicates a need to deal with more complex contradictions in a more specific manner.

The results obtained show that the created Spanish contradictions dataset is a good option for generating a language model that is able to detect contradictions in the Spanish language. This language model was capable of distinguishing between the specific type of contradiction detected. In order to create a powerful contradiction detection system in Spanish, it is necessary to extend our dataset with other types of contradictions and add specific features. This will enable us to detect, with greater precision, not only structural contradictions, but also other more complex contradictions that are possible in a real scenario for which the system is not previously trained.

Author Contributions: Conceptualization, R.S.-T. and E.S.; methodology, R.S.-T. and E.S.; software, R.S.-T.; validation, R.S.-T. and E. Saquete; formal analysis, R.S.-T. and A.B.-J.; investigation, R.S.-T.; resources, R.S.-T. and A.B.-J.; data curation, A.B.-J. and R.S.-T.; writing—original draft preparation, R.S.-T. and E.S.; writing—review and editing, A.B.-J.; visualization, R.S.-T. and A.B.-J.; supervision, E.S.; project administration, R.S.-T.; funding acquisition, E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research work has been partially funded by Generalitat Valenciana through project "SIIA: Tecnologias del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible" with grant reference PROMETEU/2018/089, by the Spanish Government through project RTI2018-094653-B-C22: "Modelang: Modeling the behavior of digital entities by Human Language Technologies", as well as being partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER) and the LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experiment could be replicated for this code (https://github.com/rsepulveda911112/ES-Contradiction-baseline, accessed on 22 March 2021). The annotation guideline and ES-Contradiction dataset are avaible in this zenodo link (https://zenodo.org/badge/latestdoi/344923645, accessed on 22 March 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
WK	World Knowledge
XNLI	Cross-lingual Natural Language Inference
RNN	Recurrent Neural Network
LSTM	Long short-term memory
GloVe	Global Vectors
CWE	Contradiction-specific Word Embedding
SVM	Support Vector Machine
SNLI	Stanford Natural Language Inference
MultiNLI	Multi-Genre Natural Language Inference
RTE	Recognizing Textual Entailment
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
NLI	Natural Language Inference
SRL	Semantic Role Labeling
Acc	Accuracy

References

1. Tudjmanand, M.; Mikelic Preradovic, N. Information Science: Science about Information. *Proc. Inf. Sci. Educ.* **2003**, *3*, 1513–1527.
2. Tsipursky, G.; Votta, F.; Roose, K.M. Fighting Fake News and Post-Truth Politics with Behavioral Science: The Pro-Truth Pledge. *Behav. Soc. Issues* **2018**, *27*, 47–70. [CrossRef]

3. de Marneffe, M.C.; Rafferty, A.N.; Manning, C.D. Finding Contradictions in Text. In *Proceedings of the ACL-08: HLT*; Association for Computational Linguistics: Columbus, OH, USA, 2008; pp. 1039–1047.
4. Dragos, V. Detection of contradictions by relation matching and uncertainty assessment. In *Procedia Computer Science*; Elsevier B.V.: Amsterdam, The Netherlands, 2017; Volume 112, pp. 71–80.
5. Lingam, V.; Bhuria, S.; Nair, M.; Gurpreetsingh, D.; Goyal, A.; Sureka, A. Deep learning for conflicting statements detection in text. *PeerJ* **2018**, *6*, e26589v1.
6. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 632–642.
7. Williams, A.; Nangia, N.; Bowman, S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*; Association for Computational Linguistics: New Orleans, Louisiana, 2018; pp. 1112–1122.
8. Chen, Q.; Zhu, X.; Ling, Z.H.; Inkpen, D.; Wei, S. Natural language inference with external knowledge. *arXiv* **2017**, arXiv:1711.04289.
9. Zhang, Z.; Wu, Y.; Zhao, H.; Li, Z.; Zhang, S.; Zhou, X.; Zhou, X. Semantics-aware BERT for language understanding. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 9628–9635. [CrossRef]
10. Conneau, A.; Lample, G.; Rinott, R.; Williams, A.; Bowman, S.R.; Schwenk, H.; Stoyanov, V. XNLI: Evaluating Cross-lingual Sentence Representations. *arXiv* **2018**, arXiv:1809.05053.
11. Canete, J.; Chaperon, G.; Fuentes, R.; Pérez, J. Spanish Pre-Trained Bert Model and Evaluation Data. PML4DC at ICLR. 2020. Available online: https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf (accessed on 22 March 2021).
12. Harabagiu, S.; Hickl, A.; Lacatusu, F. Negation, Contrast and Contradiction in Text Processing. In Proceedings of the AAAI'06 21st National Conference on Artificial Intelligence, Boston, MA, USA, 16–20 July 2006; Volume 1, pp. 755–762.
13. Lendvai, P.; Reichel, U.D. Contradiction Detection for Rumorous Claims. *arXiv* **2016**, arXiv:1611.02588.
14. Pham, M.Q.N.; Nguyen, M.L.; Shimazu, A. Using Shallow Semantic Parsing and Relation Extraction for Finding Contradiction in Text. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*; Asian Federation of Natural Language Processing: Nagoya, Japan, 2013; pp. 1017–1021.
15. Ritter, A.; Soderland, S.; Downey, D.; Etzioni, O. It's a Contradiction – no, it's not: A Case Study using Functional Relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Honolulu, HI, USA, 2008; pp. 11–20.
16. Li, L.; Qin, B.; Liu, T. Contradiction detection with contradiction-specific word embedding. *Algorithms* **2017**, *10*, 59. [CrossRef]
17. Marelli, M.; Bentivogli, L.; Baroni, M.; Bernardi, R.; Menini, S.; Zamparelli, R. SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*; Association for Computational Linguistics: Dublin, Ireland, 2014; pp. 1–8.
18. Rosemblat, G.; Fiszman, M.; Shin, D.; Kilicoglu, H. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. *J. Biomed. Inform.* **2019**, *98*, 103275. [CrossRef] [PubMed]
19. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
20. Azman, S.N.; Ishak, I.; Sharef, N.M.; Sidi, F. Towards an Enhanced Aspect-based Contradiction Detection Approach for Online Review Content. *J. Physics Conf. Ser.* **2017**, *892*, 012006. [CrossRef]
21. Sifa, R.; Pielka, M.; Ramamurthy, R.; Ladi, A.; Hillebrand, L.; Bauckhage, C. Towards Contradiction Detection in German: A Translation-Driven Approach. In Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019; pp. 2497–2505.
22. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186.
23. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
24. Takabatake, Y.; Morita, H.; Kawahara, D.; Kurohashi, S.; Higashinaka, R.; Matsuo, Y. Classification and Acquisition of Contradictory Event Pairs using Crowdsourcing. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*; Association for Computational Linguistics: Denver, CO, USA, 2015; pp. 99–107.
25. Agic, Z.; Schluter, N. Baselines and test data for cross-lingual inference. *arXiv* **2017**, arXiv:1704.05347.
26. Kuiken, J.; Schuth, A.; Spitters, M.; Marx, M. Effective Headlines of Newspaper Articles in a Digital Environment. *Digit. J.* **2017**, *5*, 1300–1314. [CrossRef]
27. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37. [CrossRef]
28. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*; European Language Resources Association (ELRA): Istanbul, Turkey, 2012; pp. 2214–2218.
29. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]

30. Zhang, Q.; Liang, S.; Lipani, A.; Ren, Z.; Yilmaz, E. From Stances' Imbalance to Their Hierarchical Representation and Detection. In *The World Wide Web Conference*; ACM: San Francisco, CA, USA, 2019; pp. 2323–2332.
31. Kang, X.; Li, B.; Yao, H.; Liang, Q.; Li, S.; Gong, J.; Li, X. Incorporating Synonym for Lexical Sememe Prediction: An Attention-Based Model. *Appl. Sci.* **2020**, *10*, 5996. [CrossRef]