DEFALAC: Automatic Detection and Generation of Fallacies Using Large Language Models Based on Deep Learning

Fermín Cruz[1], José A. Troyano[1], Fernando Enríquez[1] and F. Javier Ortega[1]

[1]University of Seville, Spain

Abstract

Our proposed research project focuses on the application of Natural Language Processing (NLP) techniques to the detection, classification and generation of fallacies. We will use large language encoder models for the detection and classification of fallacies, and encoder-decoder models for the generation of fallacies from certain premises. We will rely on manually annotated textual examples to fine-tune the pre-trained models. We intend to collect and annotate two corpus of fallacies in Spanish, one based on user conversations on the Internet, and the other based on transcripts of political debates.

Keywords

Fallacy detection, fallacy generation, large language models

1. Introduction

Logical or argumentative fallacies (hereafter fallacies) are arguments that appear to be valid but are not [1]. The invalidity of a single argument results in erroneous lines of reasoning, even though the rest of the arguments are sound. This makes the identification of fallacies a particularly interesting and necessary issue, as it would help to deactivate a truly pernicious mechanism within public debates in various contexts (political, health, economic and social, among others). The use of fallacies in these contexts, intentionally or unintentionally, contaminates debates and facilitates the spread of fake news and other disinformation phenomena. These phenomena have a great impact on society and are therefore a problem of growing interest [2], [3], [4]. Our proposed research project revolves around the application of Natural Language Processing (NLP) techniques to fallacy detection, classification and generation. This is a difficult task a priori, since the mechanisms underlying fallacies are related to semantics, discourse structure, pragmatics and knowledge of the world, among others. But this does not mean that it is unapproachable, since in recent years new tools have appeared that allow addressing tasks of similar complexity with good results.

1.1. Formal and informal fallacies

The concept of fallacy has been known since antiquity, since Aristotle first catalogued 13 logical fallacies. Since then, this catalog has only grown and become more structured. Fallacies are usually classified as formal and non-formal. Formal fallacies are those that can be detected by substituting symbols for premises and applying logical rules. Non-formal fallacies, on the other hand, are those in which this does not occur.

Formal fallacies are, for example, those that start from a logical conditional or implication ("If it is snowing, then it is cold") and apply an incorrect inference, for example through the negation of the antecedent, incorrectly inferring the negation of the consequent ("If it is not snowing, then it will not be cold"); or through the affirmation of the consequent, incorrectly inferring the affirmation of the antecedent ("It is cold, therefore, it is snowing"). As for non-formal fallacies, there are many classifications and typologies. To list some frequent non-formal fallacies, we can highlight the fallacy *ad hominem*, which consists of discrediting the person who argues, and not the argument itself. For example, this fallacy occurs when an attempt is made to discredit what is published by a journalist by arguing that he or she is paid by a certain party or company, instead of arguing against what is expressed by that person. Another very frequent non-formal fallacy is the straw man fallacy or *ad logicam* fallacy, which consists of generating a new false argument by exaggerating or caricaturing the opponent's argument, to go on to criticize this new false argument. Other very common non-formal fallacies are the fallacy of authority or *ad verecundiam*, which links the veracity of an argument to the authority of its proponent; the fallacy of false causation or *post hoc ergo propter hoc*, where it is assumed, incorrectly, that if an event occurs after another, the first is the cause of the second; or the fallacy of tradition or *ad antiquitatem*, giving veracity or justifying an argument because it is something that has always been considered so or because it is traditional; or its opposite, the argument *ad novitatem*.

To understand why fallacies are successful, we must

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations

fcruz@us.es (F. Cruz); fcruz@us.es (J. A. Troyano); fenros@us.es (F. Enríquez); javierortega@us.es (F. J. Ortega)

resort to the concept of cognitive bias. Cognitive biases are psychological shortcuts that cause a certain distortion when interpreting the information available to us at any given moment. The contribution of cognitive biases from an evolutionary point of view is to provide a mechanism for making quick decisions without performing a detailed analysis of all the available information. This speed can be vital in situations where survival itself is at stake, but it also opens the door to erroneous decision making. One of the most common cognitive biases, and one that is clearly exploited to support fallacious arguments, is confirmation bias. This bias is related to the inclination to accept as valid only that information that is aligned with our preconceived ideas. Fallacies usually take advantage of a cognitive bias to construct an argument that erroneously supports a proposition, albeit with an appearance of truthfulness. In a way, they activate the springs of the cognitive bias mechanism to appear to be correct reasoning. Fallacies may be unintentional, but in many cases they are used with the clear intention of benefiting from them.

1.2. Related NLP tasks

The task we intend to address is closely linked to other NLP tasks, including discourse analysis, argument mining and fake news detection.

Discourse analysis is a broad and complex task, which is key to applications such as summary generation, machine translation or question answering. There are different modes of discourse such as description, narration, exposition and argumentation, the latter being the one we focus on in this project. Analyzing coherence in discourse is a complex high-level task. As in many other NLP tasks, systems based on classical pipelines [5] are giving way to systems based on deep learning [6].

Argument mining consists of identifying arguments within a text and classifying them, usually as a premise or claim. Often the result is structured information that can be used as input for automatic reasoning tools aimed at detecting inconsistencies. The identified claims, often highly subjective and even controversial, are contrasted with other parts of the text identified as premises, which in principle describe more objective and proven facts, such as expert opinions, statistical values, etc., although fallacies may alter this assumption.

Another problem related to the one we intend to address is the detection of fake news or, more broadly, disinformation. Social networking services on the Internet allow users to easily redistribute, with a simple click, other users' posts. This, together with the existence of a network of connections that goes beyond the users' social network in the real world and without geographical limitations, turns social networks on the Internet into a gigantic information propagation machine. Certain content becomes viral, which means that it spreads at such a speed that it is viewed by a large number of users in a short space of time. Many of the contents that achieve a large diffusion do so because they are funny, surprising or distressing. In general, it is considered that the contents that arouse more interest (and therefore achieve more diffusion) are those that achieve some kind of emotional response in the viewers [7].

1.3. Research team experience

The research team has extensive experience in topics related to the challenges that our project proposal will pose. Since 2003 we have been working on Natural Language Processing tasks, mainly related to information extraction and automatic text classification, two techniques that will be very useful in the development of the project. We have also elaborated works in the field of social network analysis, especially with Twitter texts and also in the detection of dishonest behaviors, experiences also useful in the development of the proposal. We also have worked in the analysis of texts related to politics, one of the domains in which fallacies most often appear. Finally, we have experience in the development of resources and evaluation corpora, such as those we will need to generate for the training and evaluation of the systems we will develop during the project.

2. Objectives

The main objective of our project is to address the task of detecting, classifying and generating fallacies from natural language texts. The specific objectives are:

- To apply the latest advances in language technologies to the automatic detection and generation of argumentative fallacies in natural language, with special emphasis on the Spanish language.
- To make available to the community the necessary resources, such as corpora, models and APIs, to facilitate the development of applications that can make use of the detection, classification and generation of argumentative fallacies.
- To elaborate a catalog of argumentative fallacies and select the most appropriate ones for their approach, taking into account their frequency of occurrence in the sources studied and their logical and semantic complexity.
- To build a corpus of examples of argumentative fallacies from user conversations on the Internet, which can be used for both training and evaluation of the models.
- Build a corpus of examples of argumentative fallacies from transcripts of political debates, which

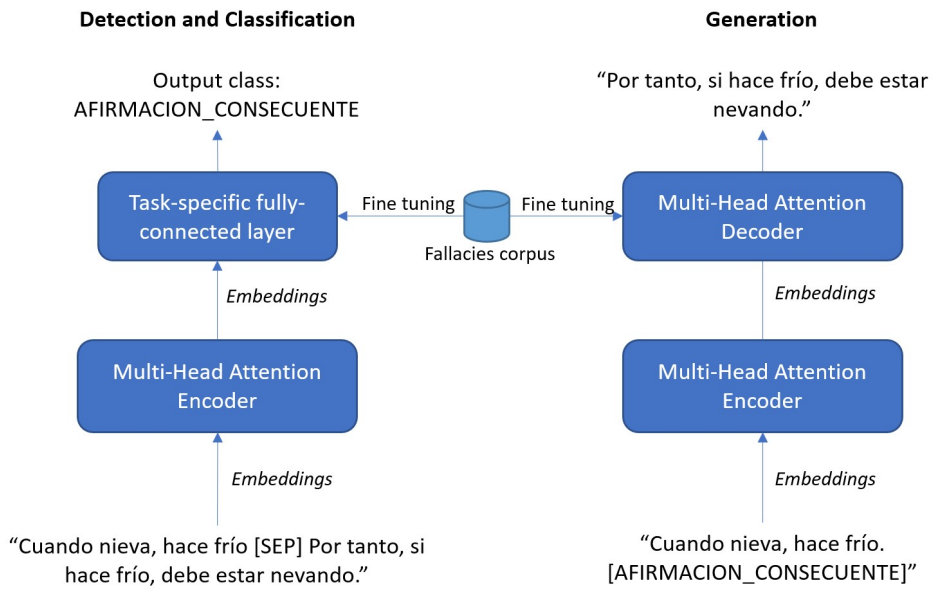

Figure 1: Abstract architecture proposal for the detection, classification and generation of argumentative fallacies based on transformers.

can be used for both training and evaluation of the models.

3. Methodology and expected results

In this section, we will summarize our methodological approach, a summary of tasks and the expected results of the project.

3.1. Methodology

For decades, NLP techniques and tools have been progressing hand in hand with concrete tasks. The tasks provide challenges and a global evaluation framework for the scientific community, while the continuous improvement of the tools makes the level of difficulty of the tasks increase continuously. One only has to review the history of any task in a competitive track that has completed several editions to see how, year after year, new aspects are incorporated that make the competition more and more demanding. It is precisely this methodological approach, based on tasks, that we want to follow during the development of the research project.

We intend to use encoder models for the detection and classification of fallacies, and encoder-decoder for the generation of fallacies from a premise. Figure 1 shows an abstract scheme of the proposed architecture we intend to implement. For fallacy detection and classification, the input to the system would be represented by a premise followed by the text in which we intend to detect the occurrence of possible fallacies. As usual in transformer-based systems, both inputs are concatenated using a special token to separate them, so that the model receives a single input sequence. The output of the encoder is produced directly by the pre-trained model that is used, so that the fine-tuning process consists of training a final fully connected layer, whose output has a neuron for each possible decision of the classifier (i.e., each type of fallacy considered plus the non-existence of any fallacy). For this purpose, we will make use of the examples available in a corpus of argumentative fallacies. In the case of fallacy generation, the input would consist of a premise plus a token indicating to the model the type of fallacy to be generated. In this case, the fine tuning consists of recalculating the decoder weights, in view of the examples available in the corpus of argumentative fallacies.

3.2. Summary of tasks

For our project, we set out to compile two corpora of fallacies, starting from two different types of texts in

Spanish: on the one hand, conversations between Internet users, such as those that can be found on Twitter or in the comments of digital newspapers or news aggregators such as *meneame.net*; on the other hand, transcripts of political debates, such as the session diaries of the Spanish Congress of Deputies, interventions in the European Parliament or transcripts of electoral debates and round tables available in different sources. On the basis of these corpora, we will address the three tasks mentioned above: fallacy detection (does a passage contain a fallacy?), fallacy classification (what type of fallacy does a passage contain?) and fallacy generation (generating fallacious arguments from a premise). The systems we will use will be based on large language models based on deep learning, also known as transformer models. These pre-trained models with large amounts of data allow fine tuning from a not too large set of examples of the task to be performed.

We will start by performing a search for publicly available annotated corpora with fallacies in English language, which will allow us to start experimentation on the tasks we want to address as soon as possible and in a preliminary way, while we develop our own corpora in Spanish. Before starting such experimentation, we will analyze the different models based on already trained transformers available that best suit our needs, both for English and Spanish, and the execution environments that best fit them. In parallel, we will elaborate a catalog of fallacies, with examples that will allow us to estimate their difficulty. Based on this catalog, we will select an initial set of fallacies to be addressed in our corpora.

As for the construction of the corpora, we will start with the one based on user conversations on the Internet, to continue with the one based on transcripts of political debates. The elaboration of both resources will be approached in a different way, due to the different nature of the sources of the texts. In the first case, we will opt for a semi-automatic approach, in which the examples of conversations with possible uses of argumentative fallacies will be obtained through the application of an algorithm; after this, a manual annotation process will be carried out, confirming or discarding the existence of fallacies in each of the examples, and selecting the type of fallacy found. In the second case, the process will be eminently manual, both in the selection of the passages and in the annotation of the observed fallacies, although we will study the use of argument mining techniques to help in the selection of the passages.

The tasks related to the training and evaluation of models for fallacy detection, classification and generation are approached sequentially. These tasks have been planned so that they can begin as soon as each of the corpora described above are developed; however, the existence of corpus of fallacies for the English language will allow us to have a preliminary experimentation task, so that when we face the experimentation with the Spanish corpus we will have as much previous knowledge as possible.

The last tasks of the project will be dedicated to the development of a software demonstrator that allows the use of the models obtained in the project and validates the proposal for its integration in final applications.

3.3. Expected results

We intend to obtain several tangible results. On the one hand, a corpus of fallacies in Spanish, and on the other hand, models for the detection, classification and generation of fallacies, ready to be used by any researcher or application developer. We also include in the objectives of our project the development of a software that demonstrates the application possibilities of these models, thus validating their practical utility.

We set out to compile two corpora of fallacies, starting from two different types of texts in Spanish: on the one hand, conversations between Internet users, such as those found on Twitter or in the comments of digital newspapers or news aggregators such as *meneame.net*; on the other hand, transcripts of political debates, such as the session diaries of the Spanish Congress of Deputies, speeches in the European Parliament or transcripts of electoral debates and round tables available in different sources. On the basis of these corpora, we will address the three tasks mentioned above: fallacy detection (does a passage contain a fallacy?), fallacy classification (what type of fallacy does a passage contain?) and fallacy generation (generating fallacious arguments from a premise).

We will start by elaborating the corpus based on user conversations on the Internet, to continue with the corpus based on transcripts of political debates. The elaboration of both resources will be approached in a different way, due to the different nature of the sources of the texts. In the first case, we will opt for a semi-automatic approach, in which the examples of conversations with possible uses of argumentative fallacies will be obtained through the users' own quotations; after this, a manual annotation process will be carried out, confirming or discarding the existence of fallacies in each of the examples, and selecting the type of fallacy found. In the case of the transcriptions of political debates, the process will be eminently manual, both in the selection of the passages and in the annotation of the observed fallacies, although we will study the use of argument mining techniques to help in the selection of the passages.

Acknowledgments

This publication is part of the project PID2021-123005 funded by MCIN/ AEI /10.13039/501100011033/ and by FEDER A way of doing Europe.

References

[1] C. W. Tindale, Fallacies and argument appraisal, Cambridge University Press, 2007.

[2] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of economic perspectives 31 (2017) 211–236.

[3] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[4] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, K. Baddour, Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter, Cureus 12 (2020).

[5] E. Cabrio, S. Tonelli, S. Villata, From discourse analysis to argumentation schemes and back: Relations and differences, in: Computational Logic in Multi-Agent Systems: 14th International Workshop, CLIMA XIV, Corunna, Spain, September 16-18, 2013. Proceedings 14, Springer, 2013, pp. 1–17.

[6] W. Song, L. Liu, Representation learning in discourse parsing: A survey, Science China Technological Sciences 63 (2020) 1921–1946.

[7] A. Dobele, A. Lindgreen, M. Beverland, J. Vanhamme, R. Van Wijk, Why pass on viral messages? because they connect emotionally, Business Horizons 50 (2007) 291–304.