Imperial College London

Evaluating Argumentative Explanations for Text Classification
Part 1: Plausibility

Participant Information Sheet

Francesca Toni (PI) and Piyawat Lertvittayakumjorn (co-I)

In this study, we will use information from you (in the form of answers to questions) for the purposes of our research. We will not collect your name or contact details or other personal data, as they are not needed for this study. Before you decide whether or not you wish to take part in this study, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully.

What is the purpose of the study?

This study aims to evaluate a new explanation method for a certain type of text classifiers. The explanation method is called AXPLR (pronounced as AX-PLORE). We want to assess whether it is more plausible and helpful for human consumption than traditional explanation methods for text classification. This information sheet is for the first half of the study aiming to assess "Plausibility" of the explanations. So, for each question, you will see a form of explanation (i.e., evidence), and you will be asked to associate the evidence to a specific class of texts. If your answer matches how the explanation method sees the evidence, we can say that the explanation is plausible.

Do I have to take part?

Participation in this study is voluntary, and you can withdraw at any time before submitting the HIT, without giving any reason and without your legal rights being affected. To withdraw, click on the "Return HIT" button, or close your browser window. If you satisfactorily complete the study, you will receive compensation for your participation, via Amazon MTurk payment system.

Technical Description - What will happen if I take part?

You will work on one of the two classification themes – either SMS Spam detection or Sentiment Analysis. Both are binary classification, having two classes of texts. For SMS Spam detection, an input text could be either *Spam* or *Not Spam*. Meanwhile, for Sentiment Analysis, an input text could be having either *Positive sentiment* or *Negative sentiment.*

For each question, you will be asked whether a given piece of evidence (a pattern, a group of phrases, or an individual phrase) is more likely to appear in which class of texts (i.e., Spam vs Not Spam for the spam detection task, or Positive vs Negative for the sentiment analysis task).

Imperial College London

Examples for the sentiment analysis task are shown below, whereas the question format will be the same for the spam detection task.

An example question regarding patterns (10 questions per HIT)

`{SENTIMENT:pos} {LEMMA:and} {SENTIMENT:pos}` A positive-sentiment word, closely followed by a form of "and", and then by a positive-sentiment word				
○ Definitely Positive	○ Positive	○ Not sure	○ Negative	○ Definitely Negative

An example question regarding groups of phrases (10 questions per HIT)

Phrase	**Your answer**				
'm returning; was return; was returned; be returned	○ Definitely Positive	○ Positive	○ Not sure	○ Negative	○ Definitely Negative

An example question regarding individual phrases (20 questions per HIT)

Phrase	**Your answer**				
'm better	○ Definitely Positive	○ Positive	○ Not sure	○ Negative	○ Definitely Negative

Your task is to select one of the five options for every question. For the spam detection task, the options are 'Definitely Spam', 'Spam', 'Not sure', 'Non-spam', and 'Definitely Non-spam'. We have provided details on how to read textual patterns as well as examples in the HIT.

Approximately, one HIT should take less than five minutes. Still, we set that a HIT must be done within one hour before it expires.

Note that all your responses will be fully anonymised. The collected (anonymised) answers will be stored in a college machine first and potentially also on a public repository such as GitHub.

What are the possible disadvantages and risks of taking part?

We are not aware of any disadvantages and risks of taking part in this study.

Reimbursement for your time

Imperial College London

If you satisfactorily complete the study, you will receive compensation for your participation, made via Amazon's payment system. The payment rates are listed below.

- For the HITs with patterns, we will pay $0.30 per HIT.
- For the HITs with groups of phrases, we will pay $0.20 per HIT.
- For the HITs with individual phrases, we will pay $0.20 per HIT.

What if something goes wrong?

If you are harmed by taking part in this research project, there are no special compensation arrangements. If you are harmed due to someone's negligence, then you may have grounds for a legal action. Regardless of this, if you wish to complain, or have any concerns about any aspect of the way you have been treated during the course of this study then you should immediately inform the Investigator (Francesca Toni, ft@ic.ac.uk). If you are still not satisfied with the response, you may contact the Imperial College Research Governance Integrity Team (rgitcoordinator@imperial.ac.uk).

What will happen to the results of the research study?

At the end of the study, we will save your answers and use them for scientific research. This will be conducted in compliance with the UK law and the recommendations and guidance published by the UK Information Commissioners Office (ICO). We will use your answers as part of a PhD thesis and we may also use them in scientific publications and also publish them (anonymously) within an open access data repository, for the purposes of scientific research by others and to ensure reproducibility.

Who is organising and funding the research?

Imperial College London will act as the main sponsor for this study. The reimbursements will be covered from the personal research fund (PRF) of the PI at the Department of Computing.

Who has reviewed the study?

This study was given approval by the Head of the Department of Computing and by the Research Governance and Integrity Team (RGIT).

Consent

Accepting the HIT amounts to consent. Specifically, by accepting the HIT, you confirm that you are 18 years old or over and you consent to carrying out this task, for your answers to be analysed for the purposes of scientific research and for your answers to be published anonymously in an open access data repository.

Contact for Further Information

Imperial College London

If you wish to obtain further information, please contact Piyawat Lertvittayakumjorn at pl1515@imperial.ac.uk.

We sincerely thank you for participating in this study.