Imperial College London

Evaluating Argumentative Explanations for Text Classification
Part 2: Tutorial and Real-time Assistance

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 second half of the study aiming to assess the usefulness of the explanations as a means to teach and support humans to perform a new task. The task we focus on in this study is *deceptive review detection*, deciding whether a given hotel review is genuine (i.e., truthful) or fake (i.e., deceptive). Previous studies reported that text classifiers (trained by machine learning algorithms) outperform humans on average in this task. So, we wonder to what extent the explanations AXPLR extracts from the classifiers are helpful to improve human performance in this task.

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, please close the Qualtrics survey window and click the "Return HIT" button in Amazon MTurk. 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 be asked to classify whether a given hotel review is truthful or deceptive. Then we will provide you a tutorial consisting of 10 questions together with explanations from the classifiers. After that, you will be asked to perform a post-test without and with assistance from the classifiers. To be specific, this survey consists of five parts.

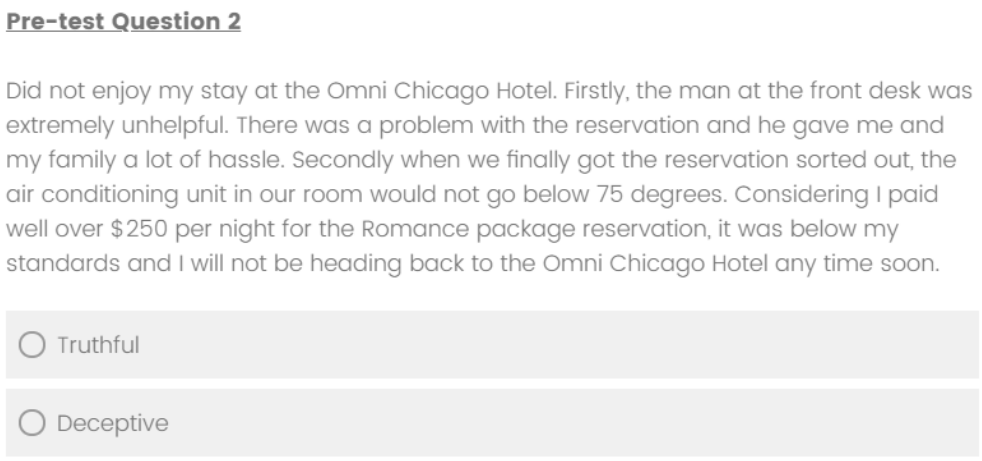
1. Attention-check questions (4 questions) -- You need to answer all the questions in this part correctly in order to proceed. This will check if you understand the task or not.

Imperial College London

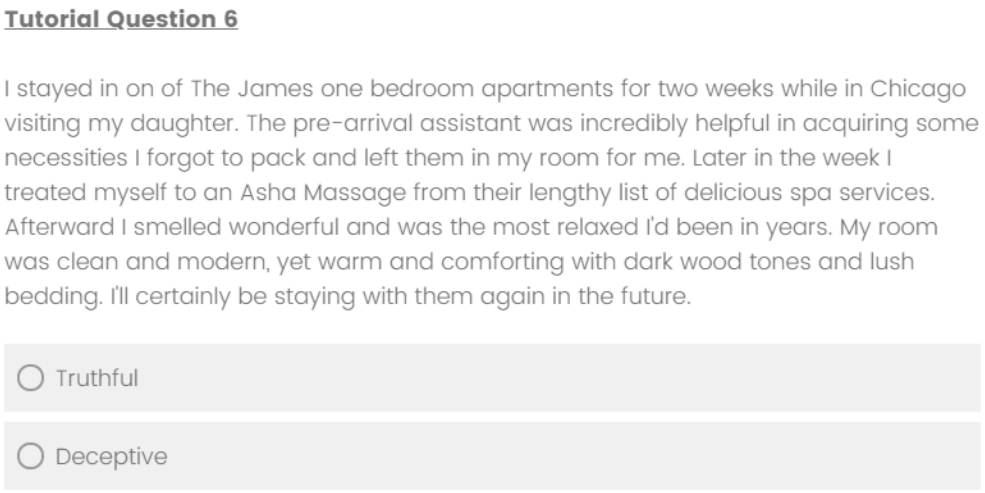
2. Pre-test (10 questions) -- For each question, you will be asked whether a given hotel review is a truthful review or a deceptive review.
3. Tutorial (10 questions) -- The format will be the same as part 2, but after each question, we will reveal to you the correct answer as well as the AI-generated prediction and explanation. (The AI prediction could be wrong sometimes.) Please use this opportunity to learn how to better detect deceptive reviews.
4. Post-test (20 questions) -- The format will be the same as part 2.
5. Additional questions (5 questions) -- You will be asked general questions about this study. Also, you may provide feedback to us in this part.

Completing all the five parts is considered as completing a HIT. One MTurker can work on only one HIT in this task.

An example of the pre-test questions



An example of tutorial question and one possible form of explanation in the tutorial phase



Imperial College London

This review is **deceptive**.

The AI also predicts **deceptive** due to the following evidence.

- Evidence for **deceptive**:

Pattern	Meaning	Match
{TYPE:smell.v}	A type of smell (v)	smelled
{TYPE:most.r}	A type of most (adv)	most
{TEXT:chicago}	The word "chicago"	Chicago
{TEXT:staying}	The word "staying"	staying
{TYPE:activity.n}	A type of activity (n)	Chicago
{TEXT:spa}	The word "spa"	spa
{TEXT:spa} {TYPE:activity.n}	The word "spa", closely followed by a type of activity (n)	spa services
{TEXT:i}	The word "i"	I

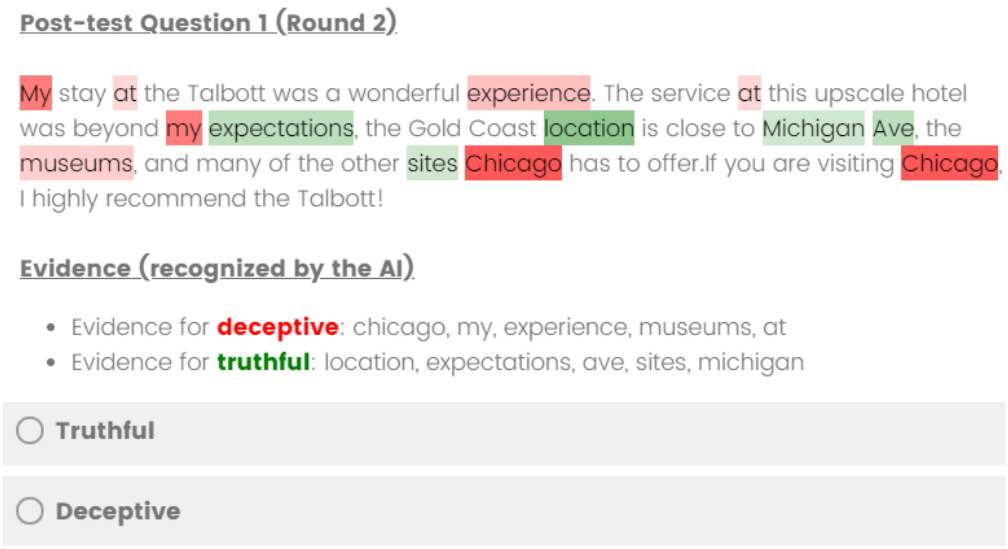
- Evidence for **truthful**:

Pattern	Meaning	Match
{TYPE:activity.n} {TEXT:i}	A type of activity (n), closely followed by the word "i"	services I
{POS:NUM}	A number	one

I stayed in on of The James one bedroom apartments for two weeks while in Chicago visiting my daughter . The pre - arrival assistant was incredibly helpful in acquiring some necessities I forgot to pack and left them in my room for me . Later in the week I treated myself to an Asha Massage from their lengthy list of delicious spa services . Afterward I smelled wonderful and was the most relaxed I 'd been in years . My room was clean and modern , yet warm and comforting with dark wood tones and lush bedding . I 'll certainly be staying with them again in the future .

An example of the post-test questions with AI assistance

Imperial College London



Approximately, one HIT should take around 30 minutes. Still, we set that a HIT must be done within four hours 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?

You may find upsetting language sometimes used in this study (appearing in negative hotel reviews).

Reimbursement for your time

If you satisfactorily complete the study, you will receive compensation for your participation, made via Amazon's payment system. For each HIT, we will give two types of reward.

1. A guaranteed reward ($2.00) will be approved for each HIT completed.
2. A bonus reward – They will be given an additional bonus reward of $0.10 for each question answered correctly (both in the pre-test and in the post-test). Therefore, the maximum bonus reward one could get is $0.10 x 30 = $3.00 per HIT. The bonus reward will be given to your MTurk account within three weeks after you complete this survey.

What if something goes wrong?

Imperial College London

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

By selecting 'I agree' for the consent question in the survey, 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

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.