Durham University

Computer Science

Summative Assignment

Module code and title	COMP4167 Natural Language Processing
Academic year	2025-26
Coursework title	RumourEval
Coursework credits	10 credits
% of module's final mark	100%
Lecturer	Noura Al Moubayed
Submission date*	06/01/2026
Estimated hours of work	20
Submission method	Ultra

Additional coursework files	https://competitions.codalab.org/competitions/16171#learn_the_details https://alt.qcri.org/semeval2017/task8/ https://alt.qcri.org/semeval2017/task8/index.php?id=data-and-tools https://aclanthology.org/S17-2006.pdf
Required submission items and formats	*Report in PDF format and a zip file of all your python code including the outputs reported in your submitted report and a README doc.*

* This is the deadline for all submissions except where an approved extension is in place. For bench tests taking place in practical sessions, the given date is the Monday of the week in which the bench tests will take place.

Late submissions received within 5 working days of the deadline will be capped at 40%. Late submissions received later than 5 days after the deadline will receive a mark of 0. It is your responsibility to check that your submission has uploaded successfully and obtain a submission receipt.

Your work must be done by yourself (or your group, if there is an assigned groupwork component) and comply with the university rules about plagiarism and collusion. Students suspected of plagiarism, either of published or unpublished sources, including the work of other students, or of collusion will be dealt with according to University guidelines (https://www.dur.ac.uk/learningandteaching.handbook/6/2/4/).

RumourEval Coursework

Students are expected to work on this coursework individually.

The task focuses on stance detection in social media conversations. Students will predict the stance of replies in a discussion thread relative to a source rumour post (tweet) using the RumourEval 2017 Subtask A dataset (SemEval 2017 Task 8). This dataset contains several rumours (source tweets) and replies annotated with the stance of the reply.

- Input: A source post (rumour) and a reply tweet.
- Output: Classify the stance of the reply towards the source post into one of four categories:
    - Support: The reply agrees with or supports the rumour.
    - Deny: The reply disagrees with or refutes the rumour.
    - Query: The reply asks for clarification or more information regarding the rumour.
    - Comment: The reply comments without expressing a clear stance (background or meta-commentary).

Students should treat the pair (source_post, reply_text) as part of the model input (fuse both fields). Students are encouraged to incorporate local thread context (previous replies) or external context.

The dataset is imbalanced, with most replies (~65–75%) labeled as Comment, while the other classes (Support, Deny, Query) occur less frequently (~5–15% each).

Students are expected to develop solutions to the stance detection task using two main methodologies:

1. Classification models: Train a supervised model (e.g., transformer-based like BERTweet) that directly predicts the four stance categories.
2. Generative / LLM-based solutions: Prompt a large language model to classify the stance of a reply given the source post.

Evaluation note: because the label distribution is skewed (Comment is dominant in the official splits), use macro-averaged F1 and per-class precision/recall/F1 as primary metrics; also show confusion matrices and class support counts.

Submit:

- Jupyter notebook (or equivalent) with code to reproduce experiments,
- README.md with commands, include random seeds, package versions, and the exact SemEval data files used.
- No more than 2000-word PDF report. Checks of word counts will be carried out on submitted work. Students are strongly advised to use Arial font size 12 for their assignments. The report word count includes all the text, including title, preface, introduction, in-text citations, quotations, footnotes, and any other item not specifically excluded below.

  Exclude diagrams, tables (including tables/lists of contents and figures), equations, bibliography/list of references.

Individual Report [100%]

Each student should separately develop their own NLP models to classify news articles into one of the four categories. Write a report (max 2000 words) on the ***challenges*** the dataset present, the ***solutions***, and your ***findings*** which will be assessed as follows:

1) **Analytics:**
   Provide analytics and insights about the task and the training and evaluation data:

   a) Compute and summarise top unigrams and bigrams for replies broken down by each stance label (S/D/Q/C). Also compare token distributions between the source text and reply text for stance (S/D/Q) vs. non-stance (Comment) classes. Provide short interpretation (what tokens signal denial, questioning, etc.). **[10%]**
   b) Apply Latent Dirichlet allocation (LDA) Run LDA separately on (i) replies labelled {S,D,Q} and (ii) Comment replies. Visualise topic word lists and word clouds and give interpretations of what you observe. **[10%]**

2) **Classification:**

   a) **4-way Classification**:

      I. Fine-tune a transformer that is appropriate for tweets (e.g., BERTweet) on the provided train/dev splits to perform 4-way classification. Input should include source+reply. Document and justify the preprocessing steps, hyperparameter choices, and imbalance mitigation chosen. Present training diagnostics (loss curve, validation macro-F1 vs epoch). [**15%**]
      II. Provide detailed analysis of the model test performance taking particular attention to the imbalance nature of the data. [**10%**]

   b) **Model Prompting**:

   I. Prompt an accessible open-source generative model to do 4-way SDQC classification. Try zero-shot and few-shot prompts. Provide exact prompt templates and the few-shot examples. [**15%**]
   II. Evaluate the classification test performance and provide recommendation of the prompt design. Report failure modes (e.g., hallucination, tendency to default to "Comment"). [**10%**].

c) **Chain of thoughts:**

i. Frame the task as a “chain of thought” task by implementing a two-stage strategy (recommended): Stage 1 classify Comment vs Non-Comment; Stage 2 for Non-Comment classify Support / Deny / Query. Compare overall 4-way performance against the direct 4-way classifier and the prompting baseline. Discuss tradeoffs (precision on minority classes, latency, error propagation). [**15%**]

3) What are the ***ethical implications*** of the dataset and your proposed solutions? What are the potential biases and future misuse cases? Provide concrete mitigation ideas. **[10%]**

4) Academic English writing, with good use of technical vocabulary, correct grammar, appropriate document structure and referencing where relevant. [**5%**]

The summative submission deadline is 14:00 on 06th of Jan 2026

The coursework aims at evaluating the students’ knowledge and their understanding of the fundamentals and advances in NLP and not their programming skills. Therefore, we ask you to implement the solutions using any Python libraries you are most comfortable with, this includes and is not limited to, PyTorch, Keras, TensorFlow, SpaCy, HuggingFace…

Jupyter notebook file should be saved along with all the produced outputs, results, and figures.

We understand that not every student has access to the same equipment and therefore this could introduce bias in model performance regardless of the of the quality of proposed solutions. Therefore, using high spec GPUs that can accelerate the performance with longer runs (e.g., epochs) will not grant the student extra marks.

FAQ

1. Can I use existing GitHub code as part of my solution?
   Yes, under the condition that you reference the code and acknowledge the authors and that the code you borrow from GitHub is not the main part of your solution.

2. Do captions count in the word limit?
   No! they do not count in the word limit. However, it is not appropriate to use diagrams or tables merely as a way of circumventing the word limit. Please be reasonable.

3. What is the logic behind choosing this topic for a coursework?
   This topic was chosen to help you strengthen your resume and career potential. Stance detection is a challenging task in NLP and has a big social impact. Talking about this task in an NLP job interview will guarantee drawing interest and questions.