School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies, Semester 2 2017

Project 2: Identifying Tweets with Adverse Drug Reactions

Due:	Stage I: 1pm (13h00 UTC+10), Wed 11 Oct 2017 Stage II: 1pm (13h00 UTC+10), Wed 18 Oct 2017
Submission materials:	Stage I: Source code, `README`, Predictions; PDF Report Stage II: Reviews (via Turnitin PeerMark)
Assessment criteria:	Method, Critical Analysis, Report Quality; Reviews
Marks:	The Project will contribute 20% of your overall mark for the subject.

Overview

The goal of this Project is to assess the effectiveness of some (supervised) Machine Learning methods on the problem of determining whether a tweet contains an ADR, and to express the knowledge that you have gained in a technical report. This aims to reinforce concepts in data mining and evaluation, and to strengthen your skills in data analysis and problem solving.

Deliverables

1. One or more programs, implemented in one or more programming languages, which generate one or more new (non-word unigram) attributes for the given tweets.
2. A README that **briefly** details how your program(s) work(s).
3. The predicted labels of the test tweets.
4. An **anonymous** technical report, of 900–1500 words, which must:
   - Give a short description of the problem and data set
   - **Briefly** summarise some relevant literature
   - Identify the newly engineered feature(s), and the rationale behind including them
   - Present the results, in terms of evaluation metric(s) and, ideally, illustrative examples
   - Contextualise the system's behaviour, based on the (admittedly incomplete) understanding from the subject materials
   - **Clearly** demonstrate some knowledge about the problem
5. In Stage II, reviews of two reports written by other students, 200–400 words each, which:
   - Briefly summarise what the author has done
   - Indicate what you think the author has done well, and why
   - Indicate what you think could have been improved, and why

Terms of Use

As part of the Terms of Use of Twitter, in using the data, you must agree to the following:

- You are strictly forbidden from re-distributing (sharing) the dataset with others, or re-using it for any purpose other than this project.
- You are strictly forbidden from re-producing messages from the collection in any publication, other than in the form of isolated examples.

Please note that the dataset is a sub-sample of actual data posted to Twitter, with almost no filtering whatsoever. Unfortunately, some of the information expressed in the tweets is undoubtedly in poor taste. We would ask you to please look beyond this to the task at hand, as much as possible.

The opinions expressed within the tweets in no way express the official views of the University of Melbourne or any of its employees; using the data in a teaching capacity does not constitute endorsement of the views expressed within. The University accepts no responsibility for offence caused by any content contained within this data.

If you object to these Terms, please contact us (`nj@unimelb.edu.au`) as soon as possible.

Assessment Criteria

Method: (15% of the marks available)
You will generate some new (non word-unigram) attribute(s); you will produce the predicted labels of the test tweets.

Critical Analysis: (45% of the marks available)
You will explain the practical behaviour of your system(s), referring to the theoretical behaviour of the Machine Learning methods where appropriate. You will support your observations with evidence, in terms of evaluation metrics, and, ideally, illustrative examples. You will derive some knowledge about the problem of identifying whether a tweet indicates an ADR.

Report Quality: (25% of the marks available)
You will produce a formal report, which is commensurate in style and structure with a (short) research paper. You must express your ideas clearly and concisely, and remain within the word limit (900-1500 words). You will include a short summary of related research.

We will post a marking rubric to indicate what we will be looking for in each of these categories when marking.

Reviews: (15% of the marks available)
You will write a review for each of two reports written by other students; you will follow the guidelines stated above.

Changes/Updates to the Project Specifications

If we require any (hopefully small-scale) changes or clarifications to the project specifications, they will be posted on the LMS. Any addendums will supersede information included in this document.

Academic Misconduct

For most people, collaboration will form a natural part of the undertaking of this project. However, it is still an individual task, and so reuse of ideas or excessive influence in algorithm choice and development will be considered cheating. We will be checking submissions for originality and will invoke the University's Academic Misconduct policy (`http://academichonesty.unimelb.edu.au/policy.html`) where inappropriate levels of collusion or plagiarism are deemed to have taken place.

Late Submission Policy

You are strongly encouraged to submit by the time and date specified above, however, if circumstances do not permit this, then the marks will be adjusted as follows:

- Each business day (or part thereof) that this project is submitted after the due date (and time) specified above, 10% will be deducted from the marks available, up until 5 business days (1 week) has passed, after which regular submissions will no longer be accepted.

Note that submitting the report late will mean that you will probably lose the opportunity for your report to participate in the reviewing process, which means that you will receive less feedback.