School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies, Semester 1 2019

Project 2: tweets r mad, or r they!?

Release: Mon 13 May 2019
Due: Report: Midnight, Tue 28 May 2019
Reviews/Reflection: 5pm, Fri 31 May 2019
Marks: The Project will contribute 20% of your overall mark for the subject; you will be assigned a mark out of 20, according to the criteria below.

Overview

The goal of this Project is to gain some knowledge about the problem of sentiment analysis on short texts. Sentiment Analysis is the process of using a Machine Learning methods to identify and categorise opinions in a piece of text in order to determine the writers attitude towards a particular product, service, topic, and so on. These can be expressed as positive, negative or neutral.

In this project you have been given real short messages ("tweets") from Twitter, under the auspices of a shared task, with the data provided by the 2017 SemEval conference. You need to use your knowledge and skills in (supervised) Machine Learning to extract some knowledge from the tweets text. Although maximizing the performance of a Machine Learning system on the given dataset is occasionally an interesting question, here we are only using the evaluation in service to help us find **knowledge**. Therefore, the main question that needs to be answered in your report is "*Can we use tweet text to help us to identify people sentiment on Twitter? If so, how? If not, why not?*"

Deliverables

1. The predicted labels of the test tweets (a separated file in TXT format)
2. An anonymous technical report, of 1100–1350 words (in PDF format), comprising 16 of the 20 marks;
3. Reviews of two papers written by your peers, each of 250-350 words (10%), comprising 3 of the 20 marks;
4. A reflection about your own paper, of 250-350 words, comprising 1 of the 20 marks.

Terms of Use

By using this data, you are becoming part of the research community — consequently, as part of your commitment to Academic Honesty, you must cite the curators of the dataset in your report:

> Rosenthal, Sara, Noura Farra, and Preslav Nakov (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. *In Proceedings of the 11th International Workshop on Semantic Evaluation* (SemEval '17). Vancouver, Canada.

If you do not cite this work in your report, you will have plagiarised these authors.

Note that the tweet collection is a sub-sample of actual data posted to Twitter, without any filtering whatsoever. As such, the opinions expressed within the documents in no way express the official views of The University of Melbourne or any of its employees, and using them in this teaching capacity does not constitute endorsement of the views expressed within. It is possible that some of the tweets are in

poor taste, or that you might find them offensive; please use your discretion when considering the data, and try to look beyond the content to the task at hand. The University of Melbourne accepts no responsibility for offence caused by any content contained within.

Report

You will submit an anonymised report (i.e. don't include your name or student ID), which should describe your approach and observations, in the machine learning algorithms you tried. And your analysis and results.

Your report should roughly use the following structure:

1. A short description of the problem, and what you aim to discover;
2. A brief summary of some published literature related to tweets sentiment analysis;
3. An overview of your classification and evaluation method(s) — you can assume that the reader is familiar with the methods discussed in this subject, and instead focus on **how and why** they relate to (and appropriate) this task and your hypothesis(es);
4. The results, in terms of the evaluation metric(s) and illustrative examples;
5. A discussion of how the results provide evidence in regards to the project main question ("*Can we use tweet text to help us to identify people sentiment on Twitter? If so, how? If not, why not?*");
6. Some conclusions about the problem of sentiment analysis of short messages from social media.

The report should consist of about 1100–1350 words. You can use tables or graphs to present the data more compactly where appropriate. You should aim to gloss over the technical details, unless they are novel or crucial to your analysis of the methods; you can assume that the reader is familiar with the methods we have discussed in this subject. **Overly long reports will be penalised**.

You should include a bibliography and citations to relevant research papers. Note that we will be looking for evidence that you have thought about the task and: have determined reasons for the performance of the methods involved; or have discerned inherent properties of the data; or can sensibly critique the problem framework. Namely, that you have acquired some **knowledge** that you can supply to the reader. A report that simply records data without corresponding analysis will not receive a strong mark.

Assessment Criteria

Report (16 marks out of 20)

A marking rubric to indicate what we will be looking for in each of these categories when marking has been posted on LMS. But in brief your report will be assessed in 3 main categories.

Method: (20% of the report mark)

You will identify a knowledge problem, and design experiments using one or more Machine Learning methods, which could plausibly be used to gain knowledge about that problem. You will describe your method(s) in a manner which would make your work reproducible from your report. You will produce the predicted labels of the test tweets.

Critical Analysis: (50% of the report mark)

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 aboutthe underlying problem of identifying a sentiment for each tweet, based on the text of it.

Report Quality: (30% of the report mark)

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 limits. You will include a short summary of related research and use related literature to support your ideas.

Reviews (4 marks out of 20)

You will write a review for each of three reports written by other students; you will follow the guidelines as stated above. 1 mark will be assigned to each completed review, and 1 mark will be assigned for overall effort. Completing the reviews is expected to take about 3–4 hours in total.

Tools

Various machine learning techniques are discussed in this subject (Naive Bayes, Decision Trees, Support Vector Machines, Association Rules, etc.); many more exist. Developing a machine learner is likely not to be a good use of your time: instead, you are strongly encouraged to make use of machine learning software and application in your attempts at this project.

One convenient framework for this is Weka: http://www.cs.waikato.ac.nz/ml/weka/.

Weka is a machine learning package with many classifiers, feature selection methods, evaluation metrics, and other machine learning concepts readily implemented and reasonably accessible. After downloading and unarchiving the packages (and compiling, if necessary), the Graphical User Interface will let you start experimenting immediately.

Weka is dauntingly large: you will probably not understand all of its functionality, options, and output based on the concepts covered in this subject. The good news is that most of it will not be necessary to be successful in this project. A good place to start is the Weka wiki (http://weka.wikispaces.com/), in particular, the primer (http://weka.wikispaces.com/Primer) and the Frequently Asked Questions. If you use Weka, please do not bombard the developers or mailing list with questions — the LMS Discussion Forum should be your first port of call.

Some people may not like Weka. Other good packages are available, most notably, scikit-learn (http://scikit-learn.org/) is quite well–regarded, if you are already familiar with Python.

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 the report 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.
- Due to the end of semester, and the inherent inconvenience caused by late submission of reviews, any submission after the reviewing system closes will incur a flat 50% penalty (i.e. 2 of the 4 marks available); reviews submitted more than 5 business days (1 week) after the deadline will not be assessed.

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