Twitter Sentiment Analysis – (theme)

IT1244 Group \_  
Arthur Chong Sze Jie, Lyu Jiapeng, Robbie Chia Yuan Ee, Tan Kai Hong Matthias

Introduction

Describe the problem that you want to solve using AI/ML, why it is important, and what AI/ML techniques that you plan to use and why. Investigate the works (2-3 recent works) that has been done with respect to the problem, what methods have been used to solve (if any) and what are their drawbacks.

Social Media has democratized the digital landscape and allowed all with internet access to express their opinions.

As of 2023, Twitter currently has 237.8 million daily active users. These users would use Twitter to express themselves, which may include posts to celebrate an achievement, rant about a negative experience, or to review a company’s product. This provides an invaluable opportunity for researchers to analyze the public’s sentiment towards certain issues. However, given the large amount of information available on social media platforms, it is extremely inefficient to determine the content of and classify each message manually. This project allows us to develop a model that can classify tweets into positive or negative classes based on the words used in it and this is often known as sentiment analysis. This project is relevant to our module as apart from being able to use simple learning machine learning methods taught in this course, we are also able to learn how to implement code to solve a Natural Language Processing task.

Currently, there are several works done to improve the efficiency in twitter sentiment analysis. Currently, there are many methods in this problem. The most basic method is the rule-based approach, where an algorithm uses a set of pre-defined rules to identify the message of a block of text. However, according to a recent comparative study on the effectiveness of this approach (<https://ieeexplore.ieee.org/abstract/document/8629140> ), the algorithm’s accuracy is heavily dependent upon the competence of the pre-defined rules. In addition, the algorithm is also less effective in interpreting contextual information, irony, and complex emotional expressions.

Another approach to sentiment extraction is an unsupervised learning technique known as word embedding. This method works through representing words as vectors and using these vectors to train classifier functions that extract the meaning of a text (<https://ieeexplore.ieee.org/abstract/document/7841054>). As an unsupervised sentiment extraction technique, it does not require pre-labelled datasets and can therefore work in a large number of different languages and contexts. However, this approach is extremely computationally expensive. In addition, it requires a large collection of text samples to generate the appropriate vectors and to train an accurate classifier function, which are not always readily available. In such cases, users are required to resort to pre-trained models, the quality of which can heavily impact the accuracy of the prediction (<https://www.sciencedirect.com/science/article/abs/pii/S095741741830558X>).

Can someone write a paragraph on why we chose the approaches we did? I am not too familiar with the Neural network and Naïve bayes, so I cant really do that

**Dataset**

Explain the dataset that you are using, what are the issues with the dataset and what analysis and processing that you did to the dataset (includes visualization and plots), etc. In addition, if you use external data to help with your project, explain the reasoning behind their use and how you used them.

In our algorithm, we were provided with a pre-labelled dataset consisting of 100,000 tweets. It contains the full message of the tweet and a number, zero or one, that identifies the attitude of the tweet as positive or negative, respectively.

The dataset is imperfect. There are many mistakes in identifying the attitude of the tweet. For instance, the positive tweet “You really gonna do a song on summer tour!? Just 1 Cant wait to hear it, comin from Scotland UK to see ya!” is labelled as zero, or negative, in our dataset. Given these errors and the large number of tweets in the dataset, it is highly unlikely that the dataset is labelled by hand. Instead, the sentiments of the messages are likely identified and labelled by another machine learning algorithm. Therefore, our results must be interpreted in the light of these: rather than reflecting the true accuracy in determining the attitude of the tweets, our accuracy score is biased by the incorrect labelling in our original dataset.

**Data Cleaning (?)**

Methods

Explain your technical approach in solving the problem that you stated in the Introduction. You will formulate the problem (mathematically if needed), explain why you chose this particular AI/ML technique. If you are using multiple methods you can explain each one of them and can justify them, if you are using AI/ML solutions that are not taught in this course, please specify them. Explain your method using figures, tables and flow-charts to explain your solution. If your method solves some limitations of previous works, then mention that as well.

[Insert Flowchart of process]

**Preparation of dataset**

The labelled dataset allows for supervised learning, where we may train a Machine Learning (ML) model to learn from it. It learns from a training set, and uses a testing set to determine the effectiveness of the model. We randomly assigned 30% of the dataset to a testing set using the *train\_test\_split* function, and the remaining 70% to a training set.

Computers are unable to read and understand textual data, hence they need to be converted into numerical data. This is done so using the Term Frequency-Inverse Document Frequency (TD-IDF) feature, which reflects the frequency and weight of each word in each tweet based on its frequency in that particular tweet and in the entire dataset. Both the train and test set have to be TF-IDF vectorised for the algorithm to read the data.

We have chosen three ML algorithms to conduct the twitter sentiment analysis – Logistic Regression, Neural Network (Long Short-Term Memory) and Naïve Bayes.

Before we implemented our algorithms, we first divided our dataset into training and testing sets. We randomly assigned 30% of the existing data to a testing set with the train\_test\_split method in the scikit-learn library and assigned the rest of the data to a training set. Then, we trained a variety of categorization algorithms with the training set and tested the resulting models against the testing set to ascertain the algorithm that is the most accurate.

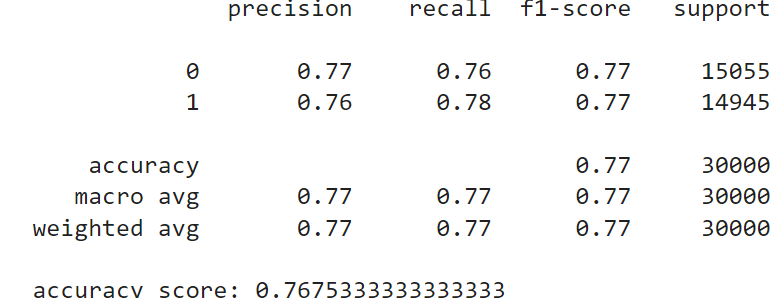
**Logistic Regression**

The first algorithm we attempted was logistic regression, which is a supervised learning algorithm that is designed to output the probability of a datapoint belonging to one of two classes. The choice was well suited to our task, which revolves around the analysis of a labelled dataset to categorize the sentiment of tweets as either positive or negative. However, because our dataset is primarily composed of strings of text, we must first convert our data into a numeric data frame that the algorithm is able to use and analyse. To accomplish this, we first removed words that we deemed unimportant, such as pronouns, possessive phrases, and prepositions. This process was expedited by filtering out the stop words provided by the National Language Toolkit in Python from the tweets. However, we were careful not to remove words that may indicate the direction of the message of the tweet, namely words such as “no”, “don’t”, “isn’t”, etc.

After removing words that are unimportant, we used the Term Frequency-Inverse Document Frequency Vectorizer function to transform the string input into a matrix of numerical Term Frequency-Inverse Document Frequency features. This matrix reflects the frequency and importance of each word in the tweet based on its frequency in the current tweet and in the entire collection of 100,000 tweets dataset. Then, we process this dataset, now fully numerical, with the Logistic Regression class that we imported from the sci-kit learns software, resulting in a model that can be evaluated using the test set.

The results can be best summarized by the confusion matrix shown in Figure (assign a number after we place it in all other graphs). Overall, as the graph reflects, the model has an accuracy of 76.75%. This - although significantly higher than guessing at random, which on average has an accuracy of 50% - needs to be more accurate for most real-world problems. Because of this, we explored the possibility of using other modelling techniques.

**Figure ()**



**Neural Network (LSTM)**

We chose LSTM because of the effectiveness of Recurrent Neural Network (RNN) in Natural Language Processing (NLP) problems, with the advantage of reducing its limitation of short-term dependencies by having memory blocks that are connected to the layers.

[Tokenizer?]

[Remainder of implementation]

[Accuracy]

**Naïve Bayes**

We chose Naïve Bayes because

[ROC AUC]

[Accuracy]

Results & Discussions

You will explain how you have evaluated the solution – how many experiments you have run, what performance metrics you have used to evaluate your model, how did you fine-tune your performance, etc. You will also report the results in tables, charts and figures. You should also list out your findings after running your experiments – explaining with evidence on why a particular model is performing poorly or well.

[Limitations]

* Seeing the entire dataset(?)

[limitation of naïve bayes is that it assumes that all features are independent, struggles with handling missing data, can be sensitive to the presence of irrelevant features.

placement. If it is necessary to include this information on the front page, use a separate footnote. The AAAI copyright notice (required) must appear first in any list of footnotes.

Abstract

All conference papers must include an abstract. The abstract must be placed at the beginning of the first column, indented ten points from the left and right margins. The title “Abstract” should appear in ten-point bold type, centered above the body of the abstract. The Abstract Head style will automatically format this heading properly.

The abstract should be set in nine-point type with ten-point leading. The Abstract Head style will automatically apply these settings. This concise summary should describe the general thesis and conclusion of your paper. A reader should be able to learn the purpose of the paper and the reason for its importance from the abstract. The abstract should be no more than two hundred words in length. (Authors who are submitting short one- or two-page extended extracts should provide a short abstract of only a sentence or so.) **Do not include references in your abstract!**

Page Numbers

Do not print any page numbers on your paper.

Text

The main body of the paper must be formatted in ten-point with twelve-point leading (line spacing). The Text style in this document is 10 point with 12-point line spacing. The first line of text after a heading should not be indented. Subsequent lines of text within the same section or subsection should be indented 10 points. The 10 point tab is automatically set in the Word template. The Text-indent style will automatically apply the indent without having to use a tab.

URLs and Word

Microsoft Word’s default setting is to underline URLs and display them in another color. This causes the URLS to turn light gray when printed, and often makes them undecipherable. **This feature must be turned off**. Please also turn off the automatic linking feature in Word. Failure to do so will result in a page reformatting fee.

Automatic Links

Do *not* use Microsoft Word’s automatic section and reference linking.

Citations

Citations within the text should include the author’s last name and year, for example (Newell 1980). Append lower-case letters to the year in cases of ambiguity. Multiple authors should be treated as follows: (Feigenbaum and Engelmore 1988) or (Ford, Hayes, and Glymour 1992). In the case of four or more authors, list only the first author, followed by et al. (Ford et al. 1997).

Extracts

Long quotations and extracts should be indented ten points from the left and right margins. The “Extract” style provides this type automatically:

This is an example of an extract or quotation. Note the indent on both sides. Quotation marks are not necessary if you offset the text in a block like this, and properly identify and cite the quotation in the text.

Footnotes

Avoid footnotes as much as possible; they interrupt the reading of the text. When essential, they should be consecutively numbered throughout with superscript Arabic numbers. Footnotes should appear at the bottom of the page, separated from the text by a blank line space and a thin, half-point rule.

Headings and Sections

When necessary, headings should be used to separate major sections of your paper. Remember, you are writing a short paper, not a lengthy book: an overabundance of headings will tend to make your paper look more like an outline than a paper.

Section headings should be twelve-point Times New Roman bold type, mixed case centered. The Section Heading style will automatically apply these settings (including the extra required spacing). Subsection headings should be eleven-point Times New Roman bold type. The Subsection Heading style will automatically apply the proper settings automatically. Do not skip a line between paragraphs. Subsubsection headings should be ten-point Times New Roman bold type, mixed case, with twelve-point leading, flush left, with three points of additional space preceding them and no additional points of leading following them. The Subsubsection heading style will apply the appropriate formatting automatically.

Section Numbers

The use of section numbers in AAAI Press papers is optional.

Section Headings

Sections should be arranged and titled as follows:

1. Main Content Sections
2. Appendices (optional)
3. Ethical Statement (optional, unnumbered)
4. Acknowledgements (optional, unnumbered)
5. References (unnumbered)

Appendices

Any appendices follow the acknowledgments, if included, or after the main body of text if no acknowledgments appear.

Ethical Statement

You can write a statement about the potential ethical impact of your work, including its broad societal implications, both positive and negative. If included, such statement must be written in an unnumbered section title *Ethical Statement*.

Acknowledgments

The acknowledgments section, if included, appears after the main body of text and is headed “Acknowledgments.” This section includes acknowledgments of help from associates and colleagues, credits to sponsoring agencies, financial support, and permission to publish. Please acknowledge other contributors, grant support, and so forth, in this section. Do not put acknowledgments in a footnote on the first page. If your grant agency requires acknowledgment of the grant on page 1, limit the footnote to the required statement, and put the remaining acknowledgments at the back. Please try to limit acknowledgments to no more than three sentences.

References

The references section should be labeled “References” and should appear at the very end of the paper (don’t end the paper with references, and then put a figure by itself on the last page). A sample list of references is given later on in these instructions. AAAI style is required for references. Poorly prepared or sloppy references reflect badly on the quality of your paper and your research. Please prepare complete and accurate citations.

Illustrations and Figures

Figures, drawings, tables, and photographs should be placed throughout the paper near the place where they are first discussed. Do not group them together at the end of the paper. If placed at the top or bottom of the paper, illustrations may run across both columns. Figures (including boxes and rules) must not intrude into the top, bottom, or side margin areas. Figures must be inserted using your page-formatting software. Number figures sequentially, for example, figure 1, and so on. There must be a minimum of 9 points between the figure and the figure caption, and an additional 12 points beneath the caption. There must also be 12 points additional space before a figure or table that does not begin at the top of a page.

The illustration number and caption should appear under the illustration. Leave some space between the figure and the caption and surrounding type; .25 inches should suffice. Captions should be presented in nine-point Times New Roman. Labels, and other text in illustrations must be at least nine-point type. The Figure Caption style will apply the appropriate formatting automatically.

Low-Resolution Bitmaps

You may not use low-resolution (such as 72 dpi) screen dumps and GIF files—these files contain so few pixels that they are always blurry, and illegible when printed. If they are color, they will become an indecipherable mess when converted to black and white. **Gif files and low-resolution jpg or png files should never be used.** The resolution of screen dumps can be increased by reducing the print size of the original file while retaining the same number of pixels. You can also enlarge files by manipulating them in software such as PhotoShop.

Using Color

The archival version of your paper will be printed in black and white and grayscale. Consequently, because conversion to grayscale can cause undesirable effects (red changes to black, yellow can disappear, and so forth), we strongly suggest you avoid placing color figures in your document. If you do include color figures, you must be mindful of readers who may happen to have trouble differentiating some colors. **Your paper must be decipherable without using color for distinction.**

**Your figures should be a minimum of 300 dpi when incorporated into your document.**



Figure 1: The First AI Magazine Cover.   
(This is an example of the required figure caption style.   
The illustration uses the required Figure stylesheet)

**You must leave a minimum of one carriage return between a figure or table and its surrounding text.**

Drawings

We suggest you use computer-drawing software (such as Adobe Illustrator or, (if unavoidable), the drawing tools in Microsoft Word) to create your illustrations. Do not use Microsoft Publisher. These illustrations will look best if all line widths are uniform (half- to two-point in size), and you do not create labels over shaded areas. Shading should be 133 lines per inch if possible. Use Times New Roman or Helvetica for all figure call-outs. **Do not use hairline width lines** — be sure that the stroke width of all lines is at least .5 pt. Zero point lines will print on a laser printer, but will completely disappear on the high-resolution devices used by our printers.

Photographs and Images

Photographs and other images should be in grayscale (color photographs will not reproduce well; for example, red tones will reproduce as black, yellow may turn to white, and so forth) and set to a minimum of 300 dpi. Do not prescreen images.

Algorithms and Listings

Algorithms are a special kind of figures. Like all illustrations, they should appear floated to the top (preferably) or bottom of the page. However, their caption should appear in **t**he header, left-justified and enclosed between horizontal lines, as shown in Algorithm 1. The algorithm body should be terminated with another horizontal line. It is up to the authors to decide whether to show line numbers or not, how to format comments, etc. We suggest placing the algorithm inside a text box to ease positioning when using MS Word.

Algorithm 1: Example Algorithm

**Input**: Your algorithm’s input

**Parameter**: Optional list of parameters

**Output**: Your algorithm’s input

1: Let t= 0.

2: **while** condition **do**

3: Do some action.

4: **if** conditional **then**

5: Perform task A.

6: **else**

7: Perform task B.

8: **end if**

9: **end while**

10: **return** solution

Listing 1:Example listing quicksort.hs

1 quicksort :: **Ord** a **=>** [a] -> [a]

2 quicksort [] = []

3 quicksort (p:xs) = (quicksort lesser) ++

[p] ++ (quicksort greater)

4 **where**

5 lesser = **filter** (< p) xs

6 greater = **filter** (>= p) xs

Listings are much like algorithms. They should also appear floated to the top (preferably) or bottom of the page. Font size in Listings must be nine-point Courier New. Listing captions should appear in the header, left-justified and enclosed between horizontal lines as shown in Listing 1. Terminate the body with another horizontal line and avoid any background color. Line numbers, if included, must appear within the text column.

References

Reference citations in the text should appear in author-year format, for example (Smith 1975). References of the same year by the same author(s) should be distinguished by small letters following the year, for example (Smith 1977c) and ordered alphabetically by title. Use a narrative citation form when referring to a paper in a narrative context. For example, say “In his paper, Michael Youngblood (2017) refers to ....” instead of (Youngblood 2017) refers to....”

All entries in the reference list must be cited in the text. In-text citations of four or more authors should be shortened to “first author et al.” For example, the university technical report reference in the sample that follows would be cited as (Vattam et. al.  2013) in the text because it contains four authors.

Generally, references include the name of the author (surname first, followed by initials only for given names) and the date, followed by a period, then the title, presented in mixed case. For multiple authors, separate two names with a comma, and three or more authors with a semicolon.  The place of publication (which is required for all book and proceedings publications) is followed by a colon, with the name of the publisher following. For journal articles and serial publications, provide the volume and issue numbers as well as the page numbers. DOIs are strongly for serial publications if they have been assigned. For conference papers, and book chapters, give inclusive page numbers. Provide the DOI if it is available. Do not use shorthand abbreviations (such as AAAI-19) — spell out the full title of the publication.

If you are citing an ephemeral or general page of a website (such as, but not limited to, the landing page of a company or product), please do not include the citation in the reference list. Instead, incorporate the URL into a footnote. All references must contain author, title, and date information.

To meet CrossRef requirements, all cited journal articles within a reference list should include a DOI if one has been assigned.

References should be listed alphabetically (by surname of the primary author or main entry) at the end of the article. Multiple references by the same author(s) should be listed in ascending chronological order with the earliest reference first (for example, Matthews 1979 precedes Matthews 1986). Information for each reference should be in the sequence illustrated by the following examples.

Dissertation or Thesis

*(Note: Include department and university):*

Clancey, W. J. 1979b. Transfer of Rule-Based Expertise through a Tutorial Dialogue. PhD dissertation, Department of Computer Science, Stanford University, Stanford, CA.

Forthcoming Book

Clancey, W. J. Forthcoming. *The Engineering of Qualitative* Models. Redwood City, CA: Addison-Wesley Publishing Company.

Preprint Server

Agrawal, A.; Batra, D.; and Parikh, D. 2016. Analyzing the Behavior of Visual Question Answering Models. arXiv preprint. arXiv:1606.07356v2 [cs.CL]. Ithaca, NY: Cornell University Library.

Published Book

Petroski, H. 1985. *To Engineer Is Human: The Role of Failure in Successful Design.* New York: St. Martin's Press.

Chapter in Published Book

Brown, J. S. 1977. Artificial Intelligence and Learning Strategies. In*Learning Strategies,* edited by J. O'Neil, 345–78. New York: Academic Press.

Forthcoming Journal Article

O'Connor, J. L. Forthcoming. Artificial Intelligence and Commonsense Reasoning. *AI Magazine*44(3).

Published Journal or Magazine Article

Cox, M. T. 2007. Perpetual Self-Aware Cognitive Agents. *AI Magazine*28(1): 32–45. doi.org/10.1609/aimag.v28i1.2027.

Paper Presented at Meeting

*(Note: Use this format only if no published proceedings appeared):*

Schoenfeld, A. H. 1981. Episodes and Executive Decisions in Mathematical Problem Solving. Paper presented at the 1981 AERA Annual Meeting. Boston, MA, September 24–30.

Zhou, S.; Suhr, A.; and Artzi, Y. 2017. Visual Reasoning with Natural Language. Paper presented at the AAAI 2017 Fall Symposium on Natural Communication for Human-Robot Collaboration. Arlington, VA, November 9–11.

Paper Presented at Meeting and Published in Proceedings

Lester, J.; Converse, S.; Kahler, S.; Barlow, T.; Stone, B.; and Bhogal, R. 1997. The Persona Effect: Affective Impact of Animated Pedagogical Agents. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.New York: Association for Computing Machinery. doi.org/10.1145/258549. 258797.

Company Technical Report

Carbonell, J. R. 1970. Mixed-Initiative Man-Computer Instructional Dialogues, Technical Report QW-19871. Marina del Rey, CA: USC/Information Sciences Institute.

Scholarly Society Technical Report

Lin, F. 2007. Finitely-Verifiable Classes of Sentences. In *Logical Formalizations of Commonsense Reasoning: Papers from the 2007 AAAI Spring Symposium*. Technical Report SS-07-05. Palo Alto, CA: AAAI Press.

University Technical Report

Vattam, S.; Klenk, M.; Molineaux, M.; and Aha, D. W. 2013. Breadth of Approaches to Goal Reasoning: A Research Survey. In *Goal Reasoning: Papers from the ACS Workshop,*edited by D. W. Aha, M. T. Cox, and H. Muñoz-Avila. Technical Report CS-TR-5029. College Park, MD: University of Maryland, Department of Computer Science.

ArXiv Paper

Bouville, M. 2008. Crime and punishment in scientific re-

search. arXiv:0803.4058.

Website or online resource

NASA. 2015. Pluto: The ’Other’ Red Planet. https://www.

nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-

12-06.

You are required to use the above AAAI reference format in your paper. If you fail to do so, your paper will be returned to you for reformatting (and a resubmission fee will apply). In addition, you must apply the buildt-in References style (which will automatically format your references in 9 point Time Roman with 10 point line spacing, and 3 additional points of space between each entry.

For the most up to date version and complete version of the AAAI reference style, please consult the *AI Magazine* Author Guidelines at

aaai.org/ojs/index.php/aimagazine/about/submissions#authorGuidelines.

Proofreading Your PDF

Please check all the pages of your PDF file. Is the page size A4? Are there any type 3, Identity-H, or CID fonts? Are all the fonts embedded? Are there any areas where equations or figures run into the margins? Did you include all your figures? Did you follow the title-case capitalization rules for your title? Did you include a copyright notice? Do any of the pages scroll slowly (because the graphics draw slowly on the page)? Are URLs underlined and in color? You will need to fix these common errors before submitting your file.

Improperly Formatted Files

In the past, AAAI has corrected improperly formatted files submitted by the authors. Unfortunately, this has become an increasingly burdensome expense that we can no longer absorb (we are charged double for papers that require reformatting). Consequently, if your file is improperly formatted, it will probably be returned to you by the outside Production agency. If that happens, you will be required to fix your file and pay a resubmission fee.

Naming Your Electronic File

We require that you name your Word source file with the last name (family name) of the first listed author so that it can easily be differentiated from other submissions. Complete file-naming instructions will be provided to you in the submission instructions.

Submitting Your Electronic Files to AAAI

Submitting your files to AAAI will be provided to you with submission instructions for your conference. Please consult this document for details on how to submit your paper.

Inquiries

If you have any questions about the preparation or submission of your paper as instructed in this document, please contact AAAI Press at the address given below. If you have technical questions about implementation of the aaai style file, please contact an expert at your site. We do not provide technical support for Word or any other software package. To avoid problems, please keep your paper simple, and do not incorporate complicated macros and style files.

AAAI Press

1900 Embarcadero Road, Suite 101

Palo Alto, California 94303-3310 USA

Telephone: (650) 328-3123

E-mail: See the submission instructions

for your particular conference or event.

Acknowledgments

The preparation of the files that implement these instructions was supported by The Live Oak Press, LLC, and AAAI Press.

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!