Rate My Tweet: Understanding Comparative Judgement in the Wild

Andy Gray
445348

Submitted to Swansea University in partial fulfilment
of the requirements for the Degree of Master of Science

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

30th September 2021

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This work is the result of my own independent study/investigations, except where otherwise stated. Other sources are clearly acknowledged by giving explicit references. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure of this work and the degree examination as a whole.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my consent for my work, if accepted, to be archived and available for reference use, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

I would like to dedicate this work to . . .

Abstract

In your abstract you should aim to summarise the core contributions of your work in the context of the problem domain. Start by outlining the domain and the problems posed within it. Discuss how the methods you focus on approach the relevant problems. You should end your abstract by concretely stating the tangible outputs and deliverables you have created in order to complete your work on this document, and whether those outputs represent and improvement or alternative approach to existing methods.

Your abstract should be a couple or so paragraphs long, and roughly approximate the order and flow you then use for structuring the main document. If a viewer has read your abstract then they should already understand at a high level what it is you have created and delivered, and whether it is better than or comparable to existing methods. If your project is driven by a research hypothesis then the reader should know what that is at a high level from this section. Reading on, little should surprise the viewer.

For paper submission of your thesis you should physically sign your name and add the date for each of the above declaration statements (black ink preferred). For digital submissions it is normally enough to simply type your name (see custard.cls), though you should sign and date them digitally using a touch or stylus input if at all possible.

Acknowledgements

This is an opportunity to acknowledge and thank those who have supported you throughout your studies. Friends and colleagues who you have studied alongside, your families, and your mentors within the department are the usual suspects.

Contents

1 Introduction **1**

1.1 Motivations 1

1.2 Overview 1

1.3 Contributions 1

2 Lit Review **3**

2.1 The Purpose of Assessment, Marking and Feedback in Education 4

2.2 Comparative Judgement 9

2.3 Related Work 11

3 Methodology **13**

3.1 Overview of Application 13

3.2 Tools 14

3.3 Ranking System 16

3.4 Data Set 17

3.5 Implementation 17

3.6 Designs 21

3.7 Risks 21

3.8 Schedule 22

3.9 Software Development Life Cycle Methodology 22

3.10 Testing 24

4 Results and Discussion **25**

4.1 Contributions 25

4.2 Future Work 25

5 Conclusions and Future Work **27**
5.1 Contributions 27
5.2 Future Work 27

Bibliography **29**

Appendices **31**

A Implementation of a Relevant Algorithm **33**

B Supplementary Data **35**

Todo list

1: Add this yourself and submit a pull request? 25
2: Add this yourself and submit a pull request? 27

Chapter 1

Introduction

1.1 Motivations

For the prior eight years, we have had involvement in some form of an educational environment. Seven of these years involve being a teacher within secondary and sixth form schools. While the focus of teaching is perceived to create lessons for students to learn and grow, we found more and more as the years went on that this wasn't the case. The focus was actually on providing reports about the students, which required data about the students from formal assessments. While having assessments to gauge the level that a student is at is an essential part of education. However, creating, marking, analysing and providing feedback for 30 students or more per class is a time-consuming task. Therefore, this assessment practice takes away the educators' time to do what is essential, creating meaningful lessons tailored for the students.

Therefore, our motivation is to create a tool for educators that will empower them to allow technology to do what it is good at and focus on what they are good at, creating and delivering lessons. To shape future generations views.

1.1.1 Objective

1.2 Overview

1.3 Contributions

The main contributions of this work can be seen as follows:

- **A LaTeX thesis template**

  Modify this document by adding additional TeX files for your top level content chapters.

- **A typesetting guide for useful primitive elements**

  Use the building blocks within this template to typeset each part of your document. Aim to use simple and reusable elements to keep your LaTeX code neat and to make your document consistently styled throughout.

- **A review of how to find and cite external resources**

  We review techniques and resources for finding and properly citing resources from the prior academic literature and from online resources.

Chapter 2

Lit Review

Education and the sharing of knowledge is a powerful tool. In fact, in our opinion the most important skill anyone can have. As a famous quote said, "give a man a fish, and he will starve, but teach him to fish, and he won't be hungry anymore". However, it wasn't until 1918 that education, as most people in England and Wales have experienced, started to come into effect [1].

Education over the years was very much about just giving the knowledge to the students from the teacher. It wasn't until 1988, under the Education Reforms Act 1988, that assessments got introduced. The introduction was through the introduction of the national curriculum in England and Wales [2].

As the curriculum got rolled out, statutory assessments got introduced to education between 1991 and 1995. Key Stage 1 first, followed by Key Stages 2 and 3, respectively [3, 4]. Only for the core subjects of English, Mathematics and Science had the assessments first introduced. The first assessments in Key Stage 1 were a range of cross-curricular tasks to be delivered in the classroom, known as standardised assessment tasks - hence the common acronym 'SATs'. However, the complexity of the use of these meant more formal assessments quickly replaced them [3, 4]. The assessments in Key Stages 2 and 3 got developed using more traditional tests.

To allow teachers to judge students' attainment, taking tests became the main assessment form in key stage 3. While assessments were the main form, educators were also able to assess their students with other means against the targets set for attainment within the national curriculum [4]. The teacher and assessment outcomes got used on a scale with key learning milestones expected at different ages. A key stage level indicated the result

for the students progress. The model was used throughout the next few years until 2005 when the role of tests in KS1 got downgraded to just being an internal support tool to teachers, and in then 2008, the government decided to remove tests in KS3 [4].

This model continued, with minor adjustments to reflect the changing content of the National Curriculum, up to 2004. From 2005, the role of the tests got downplayed at Key Stage 1, with tests being used only internally to support teacher assessment judgements [5]. Further changes came in 2008 when the government announced that testing in Key Stage 3 was to get scrapped altogether [6].

However, with a change of government party, the Conservative party taking power from the Labour party brought about new changes to how education's focuses and pedagogy methods would get conducted. In 2014 the system of attainment levels was removed, creating the educational shift of "Assessing without level" [7]. However, within schools, it was being referred to as 'life after levels'. Especially by our educational colleges and us at the time. Which was the follow up to the changes in the national curriculum in 2013 [7]. The changes within the national curriculum brought a greater focus on more traditional style GCSE academic subjects while reducing the focus on perceived technical labour style jobs. The new curriculum direction created more emphasis on the final exam outcomes at the stages of GCSE and A-Level.

2.1 The Purpose of Assessment, Marking and Feedback in Education

As we have established, assessments became a staple of the UK educational system in 1988. While the term assessments are not usually defined, the word 'assess' is typically associated with measuring, determining, evaluating, and judging [8].

While there can be multiple reasons why educators assess students, assessments aim to serve a purpose to both the teacher and the student in the process. These include: giving feedback to teachers and learners; providing motivation and encouragement; to boost the self-esteem of the pupils; a basis for communication; a method to evaluate a lesson/training method/scheme of work/ curriculum; to entertain [8]. Additionally, the assessment also creates other opportunities to rank students; a method to select and filter students, allocate students a particular pathway or educational direction, or as a way to discriminate or choose between students for a given set reason [8].

2.1.1 Traditional Methods of Assessment and Feedback

There are four main categories of assessment. These are diagnostic, formative, summative, and national assessments [8, 4]. However, it is essential to note that national assessments do not get used within everyday aspects of teaching and learning. This term is the name given to the critical exams like SATS, GCSE and ALevel exams taken nationally. Therefore we will focus on the other three main ones.

Diagnostic assessment is what gets referred to as pre-testing [8]. Educators use this technique to get a base level of knowledge of the students they have inherited. This method is good for showing the progress of attainment over time by having an initial base test. Teachers can then show how well the students have progressed over time with their improvements over the term. This base assessment also provides the teacher with crucial information - the current ability of every student's knowledge. Through knowing this current level of knowledge, teachers can adapt the coming lessons and provide suitable differentiation and scaffolding within the lessons to allow each student to succeed as much as possible. However, we also experienced, within our time as an educator, the technique getting used to create baseline narratives. Teachers were using them to show that the student's knowledge wasn't at the expected level when inherited by the teacher at meetings or performance management reviews. Therefore, being used as a counter-act measure tool by the teacher, if they find themselves being accused of letting the students' performance slip, by trying to counter-act by implying the students were not at the required level in the first place.

The second method, formative assessment, is also known as 'assessment for learning (AFL)' [8, 4]. This method has become one of the main tools for a teacher in terms of assessment and feedback. AFL allows the educator to assess the students' understanding of a topic on the fly during a lesson without a summative assessment. As a result, allowing the teacher to spend more or less time if the students do or don't get the topic, even if they planned more or less time for that topic. Therefore, ensuring that the teaching is not getting carried out for teaching sake. Thus, the emphasis is less on measurements and more on actual learning. AFL can involve using several techniques: teacher assessment - through in-class questions, marking books; to the students assessing their work called self-assessment, or peer assessment - where the students evaluate each other's work [8].

AFL has many values for teachers and students. Within Black and William's paper. 'Inside the black box [9]' discovered that AFL provides massive learning gains, especially

with the low attainer groups. Black and William found that AFL and the use of peer assessment raised motivation and self-esteem across the board, but even more so in the low attainers. With the addition of peer assessment being extra valuable to the students. This form of feedback is effective as the feedback will most likely be given back to the students in a manner that they are more familiar with, informs of language and wording. Therefore in a way that makes more sense to them and having the most impact on their learning [10, 9].

The two key ways that teachers can gain insights from AFL is in questioning and marking. Questioning, also referred to as formative questioning, aims to assess what the students in the classroom know about the current topic being discussed or taught to improve learning [8]. However, for this to be effective, students will need an appropriate 'wait time' [11]. A 'wait time' is the term used to ensure that the student, when asked a question, has to be able to formulate their thoughts and answer as the aim is not to catch them out but to gather what they currently understand. Formative questioning is also good when allowing the students to discuss amongst themselves, then answer the teacher. Therefore, allowing them to consolidate with peers to check if they understand the topic before delivering it to the teacher. A student is more likely to say they do not know than give a wrong answer and look silly in front of their peers, known as the technique 'think-pair-share'. Other effective techniques, which do not require students to discuss between themselves, are 'no-hands up', 'show-me board', 'traffic light' systems [12].

Formative marking is the term used when teachers mark students' work and provide some form of feedback, whether it be two starts and a wish or more standard approaches of providing straight-up feedback. The overall aim is to allow the teacher to see where the student is within their knowledge, gain a level of where they are at and then provide feedback of what they have done well but ultimately what they need to improve on. The proving feedback on areas to improve on are essential whether the student is at a C/4 or an A*/9. The constant feedback, no matter the students level, is as an educator always aims to ensure their students can do better. However, it is crucial that the feedback is taken on board and actioned for formative marking to be effective. Otherwise, it is more of a summative action [9, 13]. To combat this, educators would usually allow students times within a lesson, after the feedback gets given, to go back over their work and make changes to their work in a different colour.

The third method is a summative assessment, also known as 'assessment of learning (AOL) [8]. This type of assessment happens at the end of a teaching unit or topic. It gets

used to gain insights into what the students have learnt within the subject covered or the course. Its purpose is to give a student a mark, grade or ranking. Usually, this is the grade that is mainly focused on, as it is the metric that will impact the school the most in terms of league performance tables regarding GCSE and A-level results. From our experience, summative assessments are carried out regularly within schools. This assessment method tends to get used to getting a snapshot of the students of whit if a moment like, if they were to take the test now, what would they get? By seeing the results, educators can see if students need to attend intervention or if they are performing as expected or even better. With so much riding on these results, for schools and teachers performance management reviews, a lot of emphasis on put into trying to predict the final results for students. We have seen it put a lot of pressure on the teachers and the students and ultimately creates a very stressful environment, which is not the best environment for learning.

2.1.2 Why Traditional Traditional Marking and Feedback Methods are Effective

2.1.3 The Negative Aspects of Traditional Marking and Feedback Methods

While marking and feedback are essential in a classroom, they also bring about some negative aspects. As debates are happening about who formative assessment is really for [8], are these assessments for the students done to allow the students to be able to improve on their work and knowledge. Or are they more for the schools to predict actually where the students will be, come exam time. Or are they there to show external bodies, like Ofsted, that the school is being rigorous. Or are they for teachers to justify possible results based on results for their performance management reviews?

Additionally, as teachers might have had a KS4 (GCSE) class for two to three years when assessing and doing the summative assessment, the teacher might not see that student's work entirely at face value. The teacher's personal bias might jump in based on how the student has been over the year or even years. For example, if one student has been nice, well behaved and just done the required work, the teacher might provide a higher grade for that student. However, they might give a lower grade score for someone who has been a pain and misbehaved through the year. However, the second student's work might be of better quality, but it is not seen at face value and therefore not accurately marked because of the other factors.

As schools might have multiple teachers teaching a particular subject simultaneously, a process called moderation is required. Moderation aims to make sure that all work being marked and graded is all at the same level. For example, teachers A, B and C's student's work, awarded a Distinction *, are all at the exact agreed and expected quality. However, this can bring about multiple issues. One is that not all teachers might interpret the mark scheme the same as the others and therefore look for different attributes within the students' work. While moderation and standardisation aim is to find out these inconsistencies and get all the teachers on the same page regarding expectations, office politics can also hugely impact it. Imagine the scenario. Five teachers are teaching the same year group and qualification. One teacher is the lead to that subject, so, therefore, would have had all the required training from the exam boards regarding the course, another one is a regular teacher. At the same time, one is an assistant principal, another is a vice principal, and the final one is the head of the faculty. So in the whole school context, the subject lead teacher is higher in the hierarchy than the regular teacher but lower than the other three. However, in the scope of the qualification getting delivered, the lead teacher is at the top. But this can bring about the office politics we were alluding to. Some teachers who are higher up in the school system but not in the qualification scope can throw their weight around say things need to be how they have interpreted the mark scheme. Their interpretation is not always correct, but they push their view for whatever reason, bringing about a few situations. Resulting in, will the lead teacher challenge the more senior figure to say that they are wrong and the exam board expects this, or will they agree not to upset the more senior member of staff? Either way might not end well, and with the tricky world of education, the second option is the more likely choice. However, this brings about issues in regards to inconsistency with work and the awarded mark.

Another drawback to traditional marking is that the requirement of personalised feedback for students. To allow them to develop, students must have personalised areas of where they need to improve. However, in controlled assessments, teachers can give feedback, but it can not be personalised. It has to be generic, but most schools' policies require the feedback to be personalised, creating a conflict between the exam board's requirements and the school's requirements based on Ofsted's expectations. The situation makes a moral and ethical decision. They are likely to be reprimanded by the school if they do not provide the feedback but can be done for malpractice if the exam board catches them for giving the feedback.

When a summative assessment has occurred within a learning sequence, students get usually presented with a grade and feedback. This feedback and mark could be for the end of unit exams or homework, for example. While the teachers want students to focus on the feedback given to help them improve, students focus on the results and will naturally rank order themselves. The UK government has attempted to try and resolve this by removing levels in KS3. However, when KS4 focuses on the final summative assessment, their actual GCSE exams, a provided grade is hard not to offer. Therefore, it is vital to make sure that feedback is acted upon once given.

Finally, a big issue in regards to marking and providing feedback is time. It takes a long time to score a students' work and then give feedback to the students. It is also a very tedious task that a teacher might not do in one sitting. Therefore, with many potential variables in play, the marking of the points award per each exam question, for example, might not be the same. There is also a massive cognitive load that is placed upon the teacher while trying to mark.

Consequently, it is challenging to ensure that consistency and fairness are not playing a part in the marking. However, the enormous cognitive load placed upon the teacher can be very draining. It can then affect the quality of the teachers delivery within the lesson, especially with the stress aspects that get placed upon them regarding how quick the feedback needs to get returned to the students.

2.2 Comparative Judgement

2.2.1 What is Comparative Judgement

Comparative judgement is a mathematical way to determine which observation item is better than the other item also being observed compared to each other. This method was first proposed in 1927 by Louis Leon Thurstone, a psychologist, under the term "the law of comparative judgement" [14, 15]. In modern-day terminology, it gets more aptly described as a model used to obtain measurements from any pairwise comparison process. Examples of such methods are comparing the perceived intensity of physical stimuli, such as the weights of objects, and comparing the extremity of an attitude expressed within statements, such as statements about capital punishment. The measurements represent how we perceive things rather than being measurements of actual physical properties. This kind of measurement is the focus of psychometrics and psychophysics. <wikipedia>

In more technical terms, the law of comparative judgment is a mathematical representation of a discriminal process. This process involves a comparison between pairs of a collection of entities concerning multiple magnitudes of attributes. The model's theoretical basis is closely related to item response theory and the theory underlying the Rasch model. These methods are used in psychology and education to analyse data from questionnaires and tests. <wikipedia>

While comparative judgement is a technique that has been around for almost 100 years, it wasn't until the early nineties that this technique got proposed for use within an educational setting. This first proposal was by Politt and Murry [16], who conducted a study where they tested candidates on their English proficiency within Cambridge's CPE speaking exam. The judges watched 2-minute videos and judged which one out of a pair of videos they deemed better at the requested task in the exam. However, before this, in the ninety seventies and eighties, comparative judgement was presented as a more theoretical basis for educational assessments [17].

With the momentum of his findings, Politt then presented comparative judgement as a tool for exam boards to use to be able to compare the standards of A-Levels from the different exam boards, replacing the direct judgement of a script that was at the time currently being used [18]. In his papers titled, "Let's Stop Marking Exams" [19], he presents a valid argument for using comparative judgement, with the advantages it brings over some traditional types of marking.

Politt, in 2010, also presented a paper at the Association for Educational Assessment – Europe. It was about How to Assess Writing Reliably and Validly. Politt presented evidence of the extraordinarily high reliability achieved with Comparative Judgement in assessing primary school pupils' skill in first-language English writing [20].

2.2.2 The Logic Behind Comparative Judgement and What it Aims to Do

How comparative judgement works is to present two options to a marker. The marker then gets asked to pick which one of the two options they think is the better one. The marker will get presented with all possible combinations available, each time picking which one they think is the better one out of the two. An outputted score is then presented based on the method used. The original method, the Law of Comparative Judgement (LCJ), follows the formula:

S_i is the psychological scale value of stimuli i

$$S_i - S_j = x_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j},$$

Figure 2.1

However, an alternative version derived from Louis Leon Thurstone, referred to as the "Pairwise Comparison" [15], will provide an output based on the difference between the quality values is equal to the log of the odds in respect to object-A will be object-B. This formula gets represented as: $\log \text{odds}(A \text{ beats } B \mid v_a, v_b) = v_a - v_b$.

$\Pr\{X_{ji} = 1\} = \frac{e^{\delta_j - \delta_i}}{1+e^{\delta_j - \delta_i}} = \sigma(\delta_j - \delta_i)$

.

2.2.3 How effective is Comparative Judgement at Providing Feedback?

2.3 Related Work

2.3.1 Subsection all similar work

2.3.2 Comparison of similar work

Chapter 3

Methodology

In order to apply any ML and NLP to the tweet dataset, to see if we could do any information extraction and statistical analysis, we first needed to be able to generate a ranking of the ten tweets we had obtained. We sourced the tweets themed around Brexit on Twitter, and then a pipeline (see fig: 3.1) for sourcing peoples preferences of the tweets was created. The pipeline created was handled by the web app. The web app allowed the user to create an account and then compare the tweets. The resulting decision updated the ELO rating for each tweet and the more simplified traditional comparison judgment method. Each user gets only presented five different combinations, ensuring that a single tweet was only seen by the user once.

3.1 Overview of Application

The application has two main sections. The first section is a web application. This web application aims to rank the ten Twitter tweets by presenting users with two tweets and asking them which one is better. In essence, the web application is a tool to crowdsource data on peoples views based on the tweets that they get presented. The web app then

Figure 3.1: A visual representation of the processes pipline.

creates two ranking systems. One ranking system uses an ELO system, and one the users a more pairwise comparative judgement style. The pairwise comparative judgement score gets calculated by the total wins getting subtracted by the total losses.

The second section is an exploratory Python notebook looking into NLP tasks on the tweets. We carry out sentiment analysis and information extraction on the tweets to see if any patterns within the tweets match their ranking's place. For example, positive sentiment tweets getting a higher ranking with a particular theme, other than Brexit possibly showing. The ultimate aim is to create a tweet marking rubric based on the results and the information. Additionally, we will then aim to see if we can use the gained knowledge from the information extraction to see if we can train ML models to predict the tweets position within the marking grid accurately.

3.2 Tools

To create the web application and insights from the tweets, we required to use several tools. It is a requirement that we develop a full-stack web application with a user UI, an area to input the user's judgements on the tweet, store the results using a database, and extract information from the tweets using NLP techniques. Several factors within the final application needed to be satisfied for the tools to be appropriate for use.

We will be using Trello for the kanban tools. "Kanban" is the Japanese word for "visual signal" [21]. Using Kanban boards allows us to keep our work visible, this is to allow others to see what it is we are doing, and what is needed to get done. These will allow everyone to see the full picture and keep everyone on the same page.

David Anderson discovered that kanban boards get split into five components: Visual signals, columns, work-in-progress limits, a commitment point, and a delivery point [22].

Kanban teams write all their project's work items onto cards, and these are usually one per card. The kanban board gets split into columns, with each column representing an activity which composes the workflow. All the cards change between the workflow until the activity is complete. The column workflow titles can be as simple as to do, in progress and completed. However, David suggests that there should be a work in progress (WIP) limit [22]. When a column has reached the limit, of three cards, all team members get expected to focus on the cards in progress. The WIP limits are critical for exposing bottlenecks in the workflow and maximizing flow. WIP limits give an early warning sign that too much work is commissioned. Backlogs of ideas are where the

ideas of the team and the customers get placed. The moment an idea gets picked up by a team member and work begins, this gets referred to as the commitment point [22]. When the product is finished and ready for deployment, this stage is referred to as the delivery point. The overall aim of the kanban is to take a card from the commitment point to delivery point as quick as possible.

3.2.1 Programming Language

While many programming languages can handle creating a full-stack application and conducting ML, for example, Java [23], Php [24] and JavaScript [25]. We decided to use the Python language [26]. We decided upon Python due to our familiarity with it over the other main languages and its versatility. We made this decision because Python can make full-stack applications with the use of additional libraries and handle most NLP ML tasks using libraries like NLTK [27], SpaCy [28], Sci-Kit Learn [29], and TensorFlow [30].

3.2.2 Libraries

3.2.2.1 Web Application

For creating the web application, there were two main libraries available. These were Django and Flask.

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so you can focus on writing your app without needing to reinvent the wheel. It's free and open source [31].

While Flask is a small framework by most standards—small enough to be called a "micro- framework," and small enough that once you become familiar with it, you will likely be able to read and understand all of its source code [32].

Flask has three main dependencies. The routing, debugging, and Web Server Gateway Interface (WSGI) subsystems come from Werkzeug; the template support is provided by Jinja2; and the command-line integration comes from Click. These dependencies are all authored by Armin Ronacher, the author of Flask [32].

Flask has no native support for accessing databases, validating web forms, authenticating users, or other high-level tasks. These and many other key services most web applications need are available through extensions that integrate with the core pack-

ages. As a developer, you have the power to cherry-pick the extensions that work best for your project, or even write your own if you feel inclined to. This is in contrast with a larger framework, where most choices have been made for you and are hard or sometimes impossible to change [32].

After experimenting with the two frameworks, we decided upon Flask. Flask got decided upon because of the short time frame to put the project together. Additionally, the lightweight nature of the framework also played a fact. As this will be just an initial prototype, Django's other requirements would be unessential additionals to the project. Therefore, taking focus away from what we believe is the main focus.

3.2.2.2 NLP Tasks

3.2.3 IDE

While many great IDEs are available like Pycharm, Jupyter Lab, Atom and Sublime, we decided to use VS Code. The decision behind this was that it allowed us to explore code within interactive python notebooks (ipynb) and standard python scripts. Additionally, it allowed us to create HTML, CSS, and Javascript files within the same IDE.

3.3 Ranking System

As discussed in the literature review, along with a more traditional pairwise comparative judgment algorithm, we could choose either an ELO or Glicko system. While each has advantages and disadvantages, we decided to use the ELO system. We decided to use this system as we felt it would be more robust for how we intend to be calculating the tweet scores, as we will be taking random pairings of tweets that will only be seen once by the user. Only seeing the tweet appear once removes any opportunity for a user to underrate a tweet because it has been seen multiple times and not lose its impact on the user.

Due to this reason, the ELO system, with its probability aspect to the scoring, helped determine outcomes on potential unseen tweet combos. While not considering if a tweet gets seen more than any others, this would have a massive impact on the comparative judgement pairwise comparison method.

$$\text{Prob A Wins} = 1/1 + 10^{(B-A/400)}$$

Figure 3.2: To calculate the expected score for a tweet.

$$\text{new score} = rating + 32 * \text{score} - \text{expected score}$$

Figure 3.3: A visual representation of the web apps navigation.

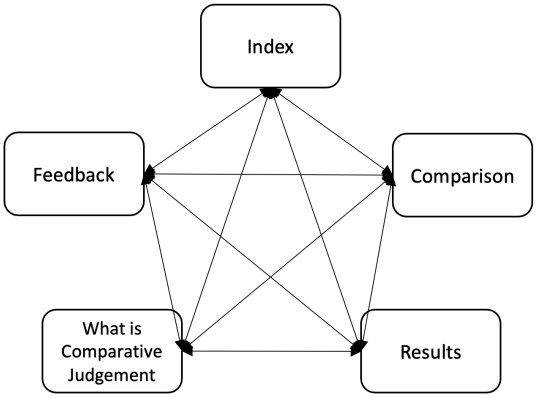

Figure 3.4: A visual representation of the web apps navigation.

3.4 Data Set

3.4.1 Data Capture Method

3.4.2 Pre-Processing

3.5 Implementation

The web application got implemented using the Python web library Flask. The web application used several industry-standard tools, for example, HTML, CSS, JavaScript, Bootstrap and dynamic content. The HTML, CSS, Bootstrap and JavaScipt was used to handle the application's front end. The web application had a mesh style navigation system (see fig: 3.4). However, when the user was on the compare page, this would push to itself and update the users content based on what they had next in their comparison list.

Additional tools like Google's Firebase was used to handle user authentication and store the web app's content in their real-time databases. The real-time databases are a NoSQL document notation database that updates in real-time.

A requirement of the app is for the user to be able to create an account. The account sign-up only requires an email and will generate all the additional requirements for the other parts of the app to work in the background. They are linking all the results for these comparisons to the user's ID. At the point of sign-up, a user position within the comparison cycle gets generated, a random selection of tweets to get compared against will be generated. The logic behind the sampling is that a user will only see a single tweet once. Therefore making sure that the user sees these tweets for the first time, every time, making it more of a fair comparison.

Heroku handled the hosting of the web app. Heroku is a free-to-use web hosting provider. However, with it being a free-to-use service, it did bring about some undesirable aspects, mainly the website's slow loading time.

As previously mentioned, a user will have a random sample of the tweets, which will have a unique pairing. Therefore ensuring that a user will only see one tweet within the pairing once, to make the tweet's joke not lose its impact as the second or third time a user sees the same tweet, it naturally would lose its edge. Hence, each user will have their own predetermined set of comparisons at the point of sign-up but will one see, for example, tweet one once. As we mentioned, this was to keep the tweets fresh for the user and make them more likely to complete all the comparisons. Otherwise, if the user had to see all unique comparisons, they would have to see 45 different combinations in total just for ten different tweets. So if we put this into the context of a teacher, who would usually have 30 students in a class, several teachers will have to see 435 different combinations, which is just for one class. When this gets factored in, we are looking at around 11175 for 150 different students.

The app will query the database and look for the user's current position when presenting the tweets. Based on their position, the tweet combinations then get checked for that according to the round. The tweet ids are then queried against the tweets' content and then presented to the web page. The user gets expected to select a tweet that they find funnier and then provide an opportunity to justify their choice, which is optional.

When the user presses the "Vote!" button, this saves the results to the database, updating the two result systems and the user's position. The process will save which tweet won and lost and update the ELO ranking and the standard ranking. The standard ranking gets calculated by taking how many times a tweet has won minus the number it has lost. The implementation of the standard ranking system is to try to implement a more

$$ProbAWins = 1/1 + 10^{(B-A/400)}$$

Figure 3.5: To calculate the expected score for a tweet.
new score $= rating + 32*$ score $-$expected score

Figure 3.6: A visual representation of the web apps navigation.

traditional comparative judgement ranking system. In contrast, the ELO system is using a more traditional approach (see fig: 3.6) Which gets updated after every comparison. The implementation of the two systems allows us to see if the ELO or more standard version of CJ is the more effective one or if they naturally mirror each other.

This process gets repeated until the user has completed all five comparisons.

3.6 Designs

3.6.1 Home Page

3.6.2 Comparison Page

3.6.3 What is Comparative Judgement Page

3.6.4 Results Page

3.6.5 Feedback Page

3.7 Risks

*S= Severity, L = Likelihood, D= Detection

Risk	S	L	D	RPN	Mitigation
The application is not user friendly.	6	3	2	36	Through user testing, to gain feedback and review.
Application does not meet expectation of the user.	6	3	3	54	User testing must be carried out and feedback taken to adapt the app.
Application has foundation bugs which effects performance.	9	3	6	162	Making sure app that the app is carrying out the core requirements correctly is essential.
Loss of Data/ Application.	8	3	7	168	To make sure that solution is back up by using services like GitHub and other back-up solutions.
More time needed to complete required tasks.	7	4	6	168	Any additional tasks that are not essentially required will have to get discarded.
Not enough time to learn required libraries to highest level.	4	4	6	96	Make sure that NLP and Flask is learnt well enough to be able to put the main concept together.
Inability to incorporate NLP into the research.	6	6	7	252	Make sure that the ELO and Comparative Judgement rankings are carried out correctly.
Under estimation of the project's complexity.	7	5	3	105	Define the projects scope clearly and learn required skills needed to complete the task.
Unrealistic time estimations.	7	4	1	28	Essential that all times requirements are followed. If falling behind, then escalation to project supervisor is required and time management redone.
Failure to follow the project's planed methodology.	6	3	1	18	Ensure requirements to methodology are clear.

3.8 Schedule

3.9 Software Development Life Cycle Methodology

Project management is crucial for any task that is about to be carried out, even more so for software development. As a famous Benjamin Franklin quote says, "Failing to plan is planning to fail" [33]. With this in mind, we must decide on the suitable project planning method that compliments our initial software design. From the waterfall method to Rapid application development (RAD) or the more modern methods of agile development, there are many methods that we could choose. We will explain the different methods we could use and what would be best for our solution and intended development method.

The profession of the software developer has existed since the first computers, but the practices and methods for developing software have evolved over timer [34]. The approaches have developed over the years to adapt to the ever-changing landscape of software development. The methods, known as software development life cycles (SDLC), vary in approach but fundamentally share the same goal. The main aims of the SDLC are to break the development up into stages. However, what changes with different SDLC is how these stages get carried out. The different stages are planning, requirements, designing and prototyping, software development, testing, deployment, operations, and maintenance [34].

The first stage, planning, involves resource allocation, capacity planning, project scheduling, cost estimation, and provisioning [34]. The primary outcome of this stage is to have an overall plan of what we have and what we will need to complete our goal within the constraints like costs and times allowed. The second stage, requirements, is where Subject Matter Experts (SMEs.) guide on what would be needed to carry out the stakeholders' requirements [34]. The third stage, design and prototyping, is where the software architects and developers begin to design the software. The outcome of this stage would be documentation on the intended design patterns and design wireframes of the intended final software. The fourth stage, development, is where the software starts to get made based on the decisions made in design and prototyping, following the chosen methodology. The outcome will be testable, tangible software. The fifth stage, testing, is considered the most crucial stage [34]. It is essential to do all the code quality checking, unit testing, integration testing, performance testing and security testing. The sixth but by no means the final stage is deployment. This stage is when the code is ready

to be shipped to the client or uploaded to the required app stores. However, the final stage is operations and maintenance. This stage is about ensuring that the software is getting used as it should and that any bugs that did not initially get picked up in testing are correct and removed from the software.

The waterfall method is a model where each section needs to be completed before moving onto the next stage, like a waterfall flowing down. For example, before we can start analysing the requirements, we need to complete the planning stage. Following the seven critical stages of SDLC, one after the other.

Like all models, they have their advantages and disadvantages. Advantages that this model has is that it is easy to use and follow, and by the way it is all set up, every stage will get finished before the next stage starts. The waterfall method also allows for the project to be easily managed, resulting in easier documentation [**?**]. However, some of the disadvantages are that it is not very useful if the requirements are not very clear at the beginning. Another disadvantage is that once we have moved to the next stage, it is tough to go back to a previous stage to make any changes which therefore creates higher risks to development and has less flexibility [**?**]. The model is best when changes in the project are stable, and the project is small, with the project requirements are clearly defined.

The overall aim of RAD is to create software projects with higher quality and faster by gathering requirements through workshops or focus groups. Then prototyping the product and then using reiterative user testing of designs early. RAD is the best model for when we need something created quickly and have a pool of users available to test prototypes. However, this approach can be costlym [**?**].

The Spiral Model is an SDLC methodology that aids in choosing the optimal process model. It combines aspects of the incremental build model, waterfall model and prototyping model but is different by a set of six invariant characteristics [35]. The Spiral Model main focus is on risk awareness and management. The risk-driven approach of the spiral model ensures the team is highly flexible within its approach and highly aware of the challenges they can expect down the road. The spiral model shines when stakes are highest, and significant setbacks are not an option [35].

The Agile methodology is a process by which a team can manage a project, which gets achieved by breaking up the project into several stages. It required constant collaboration with stakeholders, which leads to continuous iterations of improvement. In essence, Agile development is not a set methodology more of a manifesto aiming to uncover

better ways to develop software. "Individuals and interactions over processes and tools. Working software over comprehensive documentation. Customer collaboration over contract negotiation. Responding to change over following a plan [**?**]."

The project's requirements have features that lend themselves well to the waterfall methodology. However, we would like to have an element of agile methodology within the development due to the application intending to get created in a modular way. Using the waterfall method will allow us to have a clear plan and requirements of what is needed, but by using the agile method, we can rotate between the software development and testing stages.

3.10 Testing

The web application was the part of the implementation that required rigorous testing. The testing was because the web app was the bit that users would be interacting with the study. Therefore, we needed to ensure the app was to a high standard not to detract away from the users' experience and solely focus on the application purpose, which is to select which tweet they think is funnier.

We conducted multiple in-house testing using an internal server's localhost to ensure that the app was suitable. Additionally, we allowed a small number of users to test out the application. Once we were happy with the feedback, the application's data got reset and published to potential users.

Chapter 4

Results and Discussion

In this document we have demonstrated the use of a LaTeX thesis template which can produce a professional looking academic document.

4.1 Contributions

The main contributions of this work can be summarised as follows:

- **A LaTeX thesis template**

  Modify this document by adding additional top level content chapters. These descriptions should take a more retrospective tone as you include summary of performance or viability.

- **A typesetting guide for useful primitive elements**

  Use the building blocks within this template to typeset each part of your document. Aim to use simple and reusable elements to keep your document neat and consistently styled throughout.

- **A review of how to find and cite external resources**

  We review techniques and resources for finding and properly citing resources from the prior academic literature and from online resources.

4.2 Future Work

Future editions of this template may include additional references to Futurama.

1: Add this yourself and submit a pull request?

Chapter 5

Conclusions and Future Work

In this document we have demonstrated the use of a LaTeX thesis template which can produce a professional looking academic document.

5.1 Contributions

The main contributions of this work can be summarised as follows:

- **A LaTeX thesis template**

  Modify this document by adding additional top level content chapters. These descriptions should take a more retrospective tone as you include summary of performance or viability.

- **A typesetting guide for useful primitive elements**

  Use the building blocks within this template to typeset each part of your document. Aim to use simple and reusable elements to keep your document neat and consistently styled throughout.

- **A review of how to find and cite external resources**

  We review techniques and resources for finding and properly citing resources from the prior academic literature and from online resources.

5.2 Future Work

Future editions of this template may include additional references to Futurama.

2: Add this yourself and submit a pull request?

Bibliography

[1] UK Public General Acts, "Education act 1918," 1918.

[2] ——, "Education act 1988," 1988.

[3] D. Hutchison and I. Schagen, *How reliable is National Curriculum assessment?* NFER, 1994.

[4] J. Dillon and M. Maguire, *Becoming a teacher: Issues in secondary education*. McGraw-Hill Education (UK), 2011.

[5] BBC News. (2004) Primary school tests toned down. [Online]. Available: http://news.bbc.co.uk/1/hi/education/3656244.stm

[6] ——. (2008) Tests scrapped for 14-year-olds. [Online]. Available: http://news.bbc.co.uk/1/hi/education/7669254.stm

[7] Department for Education. (2013) Assessing without levels. [Online]. Available: https://webarchive.nationalarchives.gov.uk/ukgwa/20130802141012/https://www.education.gov.uk/schools/teachingandlearning/curriculum/nationalcurriculum2014/a00225864/assessing-without-levels

[8] J. Wellington, *Secondary education: The key concepts*. Routledge, 2007.

[9] P. Black and D. William, "Inside the black box: Raising standards through classroom assessment. phi delta kappam," 1998.

[10] H. Torrance and J. Pryor, *Investigating formative assessment: Teaching, learning and assessment in the classroom*. McGraw-Hill Education (UK), 1998.

[11] P. Black and C. Harrison, "Feedback in questioning and marking: The science teacher's role in formative assessment," *School science review*, vol. 82, no. 301, pp. 55–61, 2001.

[12] OECD. (2005) Formative assessment: Improving learning in secondary classrooms. [Online]. Available: https://www.oecd.org/education/ceri/35661078.pdf

[13] D. William, "National curriculum assessment arrangements," *British Journal for Curriculum and Assessment*, vol. 1, pp. 8–12, 1990.

[14] L. L. Thurstone, "Psychophysical analysis," *The American journal of psychology*, vol. 38, no. 3, pp. 368–389, 1927.

[15] ——, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[16] A. Pollitt and N. L. Murray, "What raters really pay attention to," *Studies in language testing*, vol. 3, pp. 74–91, 1996.

[17] D. Andrich, "A rating formulation for ordered response categories," *Psychometrika*, vol. 43, no. 4, pp. 561–573, 1978.

[18] P. Newton, J.-A. Baird, H. P. Harvey Goldstein, and P. Tymms, "Paired comparison methods," 2007.

[19] A. Pollitt, "Let's stop marking exams," 01 2004.

[20] ——, "Abolishing marksism and rescuing validity," *International Association for Educational Assessment, Brisbane, Australia. http://www. iaea. info/documents/paper_4d527d4e. pdf*, 2009.

[21] M. Rehkopf, "What is a kanban board?" Retrieved April 29, 2020 from: https://www.atlassian.com/agile/kanban/boards.

[22] D. J. Anderson, *Kanban: successful evolutionary change for your technology business*. Blue Hole Press, 2010.

[23] K. Arnold, J. Gosling, and D. Holmes, *The Java programming language*. Addison Wesley Professional, 2005.

[24] S. S. Bakken, Z. Suraski, and E. Schmid, *PHP Manual: Volume 1*. iUniverse, Incorporated, 2000.

[25] D. Flanagan, *JavaScript: the definitive guide*. " O'Reilly Media, Inc.", 2006.

[26] Python Core Team, *Python: A dynamic, open source programming language*, Python Software Foundation, Vienna, Austria, 2020. [Online]. Available: https://www.python.org/

[27] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[28] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[31] Django. (2021) Meet django. [Online]. Available: https://www.djangoproject.com/

[32] M. Grinberg, *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.

[33] N. Raje, "Failing to plan is planning to fail." Retrieved April 29, 2020 from: https://www.batimes.com/articles/failing-to-plan-is-planning-to-fail.html.

[34] D. Swersky, "The sdlc: 7 phases, popular models, benefits & more [2019]." Retrieved April 29, 2020 from: https://raygun.com/blog/software-development-life-cycle/.

[35] A. Powell-Morse, "Spiral model: Software development for critical projects." Retrieved April 29, 2020 from: https://airbrake.io/blog/sdlc/spiral-model.

Appendix A

Implementation of a Relevant Algorithm

```
#include <stdio.h>

int main(int argc, char *argv[]) {
  printf("Hello world.\n");
  return 0;
}
```

Listing A.1: An implementation of an important algorithm from our work.

Appendix B

Supplementary Data

The results of large ablative studies can often take up a lot of space, even with neat visualisation and formatting. Consider putting full results in an appendix chapter and showing excerpts of interesting results in your chapters with detailed analysis. You can use labels and references to refer the reader here for the full data.