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

Todo list

Chapter 1

Introduction

We have set out to create a tool that can simulate a small scale comparative judgement experiment on what users think about tweets getting compared against each other. This experiment is in light of our stakeholder getting commissioned by the Welsh government to implement a comparative judgement system nationally for all schools in Wales. Comparative judgement is a technique that has been around for almost 100 years. However, while the process can improve results and reduce cognitive loads for teachers and markers, especially at the scale that the stakeholder's implementation will have to work at, it can still require many combinations to be marked and compared. For this experiment, we decided to use tweets based on Brexit.

Therefore, we have created a tool that allows users to see a sub-sample of the combinations. Once the users have viewed the varieties, an overall ranking of the results will get created. Two methods got implemented, a more traditional comparative judgement method and an Elo style ranking.

We then aimed to use NLP techniques to see any insights we could find within the tweets. We intended to extract information on the tweets to see if we could find patterns that would give us insights into what might have impacted the tweets final scores.

The study got broken up into two parts. Part one was a web app to gather user's views on the tweets, and the second part was exploring NLP techniques within a Jupyter Notebook. With our aim to see if we can generate any feedback about the tweet.

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.

1.2 Existing Liturature

1.3 New Insights

While the comparative judgement technique has many great features, we believe that the concept can still improve. We believe this is especially the case when the comparative judgment system gets expected to get done at a national scale. We believe this because the traditional method would expect all unique pairings to get compared. Additionally, the adaptive comparative judgement that most other systems have adopted still requires time and effort even when the number of individual student work is only around thirty. Therefore, it would be tough to do when needed to get scaled up to a national level. That is why we believe a different ranking system, like an Elo system, could replace the adaptive comparative judgement process and have a more crowed sourced approach. Therefore, reducing cognitive load and the time cost it would take for people to partake.

Furthermore, the current implementations do not provide any meaningful feedback to the students or educators about what makes a piece of work better than the other. Therefore, we think we can look into NLP techniques that can provide some form of feedback. To

see if this can become something more meaningful and give some insights. Marking and giving feedback is a crucial role for all educators and the students receiving the feedback.

1.4 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.

1.5 Results Overview

1.6 Overview

We will first look into the background, explaining the need education has for marking, allowing educators to rank students' work, and providing feedback to students to enable them to reflect and improve. We will then look into what comparative judgement is and its different iterations. Additionally, we look into different ranking systems, with both coming from the chess world but get currently implemented in all other scenarios, like e-Sports. We then look into what Natural Language Processing (NLP) is and some techniques to help achieve what we aim to achieve within our implementation. Then finally for this section will look at other applications that aim to implement comparative judgment within them. We will then look at our methodology, explaining the tools and design approaches we decided to use. We then look at the results we found and a discussion around these. We then finish with a conclusion and suggested further work for this project.

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 better. The marker will get presented with all possible combinations available, each picking which one they think is better out of the two. An outputted score is then presented based on the method used, providing a preference order of observations.

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:

$$\log \text{odds}(A \text{ beats } B \mid v_a, v_b) = v_a - v_b \, .$$

Pairwise comparison generally is any process of comparing entities in pairs to judge which of each entity is preferred or has a greater amount of some quantitative property, or whether or not the two entities are identical. The pairwise comparison method get used in the scientific study of preferences, attitudes, voting systems, social choice, public choice, requirements engineering and multiagent AI systems.

Within an educational setting, a different approach of comparative judgement has been proposed. This new adaptation gets referred to as adaptive comparative judgement (ACJ) [21]. It is also the same as the pairwise comparison, just with a different name. ACJ is very similar to the core concept of comparative judgement, as it asks a marker to rate which work is better. However, in this version, the 'scores', the model parameter for each object, get re-estimated after each 'round' of judgements. Resulting in each piece of work being judged one more time on average. During the next round, each piece of work is compared only to another whose is currently estimated to have a similar score. Therefore, comparing each piece of work with a similar score results in an increased amount of statistical information from each judgment to produce the final ranking. As a result, the estimation procedure is more efficient than random pairing or any other pre-determined pairing system like those used in classical comparative judgement applications [21].

2.2.3 What does ACJ aim to achieve and How reliable is it

Multiple studies have shown that ACJ achieves exceptionally high levels of reliability, often considerably higher than the traditional method of marking. It, therefore, offers a radical alternative to the pursuit of reliability through detailed marking schemes [21].

ACJ software estimates a 'measure' for each piece of work getting compared, known as a 'script'), and an associated standard error. The process requires several metrics to be measured. These are the true SD, SSR and the index G [22].

The 'true SD' gets calculated for the script by using the formula:

$$(TrueSD)^2 = (ObservedSD)^2 \text{ - } MSE$$

The MSE represents the mean squared standard error across the scripts.

The SSR gets defined like reliability coefficients in traditional test theory, as the ratio of true variance to observed variance with the formula:

$$SSR = (TrueSD)^2/(ObservedSD)^2 \ .$$

Sometimes another separation index G is calculated. Index G represents the ratio of the 'true' spread of the measures to their average error. The formula is:

$$G = (TrueSD)/RMSE$$

The RMSE is the square root of the MSE. Leading to the SSR, as an alternative, to be calculated as:

$$SSR = G^2/(1 + G^2)$$

Studies have found that ACJ has high reliability, even compared to the final results when work is marked more traditional. However, frustration has been prevalent when markers have had to review repetitive work [23]. Additionally, frustration also gets created by the lack of students being able to challenge the final results [23].

When we look at fig: 2.1, we can see that these studies have produced a high SSR score. However, a lot of the studies have used a high resource count to complete the different studies. For example, Pollitt 2012 studies used 54 judges to mark 1000 pieces of scripts, which resulted in 8161 different comparisons getting seen and 16 rounds occurring. In comparison, Whitehouse & Pollit (2012) had 564 scripts to compare and 23 judges. This study took 12 - 13 rounds to get a high SSR score. Therefore, we can see that while ACJ can help with teacher workload in removing a cognitive overload, it takes to create additional workload in the sheer amount of rounds required to get a reliable SSR score.

Additionally, a number of the studies have used 20 - 100 different judges, which is more than most teachers within a single department. Therefore, making it hard for us to see how within a typical set-up of a school. As in, does the requirement needed to produce an accurate judgment outweigh the reduced cognitive load.

Studies have found that ACJ has high reliability, even compared to the final results when work is marked more traditional. However, frustration has been prevalent when markers have had to review repetitive work [23]. Studies have found that ACJ has high reliability, even compared to the final results when work is marked more traditional. However, frustration has been prevalent when markers have had to review repetitive

Study	Adaptive?	What was judged	#scripts	#judges	#comps	%max	#rounds	Av. # comps per script	SSR
Kimbell et al (2009)	Yes	Design & Tech. portfolios	352	28	3067	4.96%		14 or 20 bimodal	0.95
Heldsinger & Humphry (2010)	No	Y1-Y7 narrative texts	30	20	~2000?			~69	0.98
Pollitt (2012)	Yes	2 English essays (9-11 year olds)	1000	54	8161	1.6%	16	~16	0.96
Pollitt (2012)	Yes	English critical writing	110	4	(495)	(8.3%)	9	~9	0.93
Whitehouse & Pollitt (2012)	Yes	15-mark Geography essay	564	23	3519	2.2%	(12-13)	~12.5	0.97
Jones & Alcock (2014)	Yes	Maths question, by peers	168	100,93	1217	8.7%	N/A?	~14.5	0.73 0.86
Jones & Alcock (2014)	Yes	Maths question, by experts	168	11,11	1217	8.7%	N/A?	~14.5	0.93 0.89
Jones & Alcock (2014)	Yes	Maths question, by novices	168	9	1217	8.7%	N/A?	~14.5	0.97
Newhouse (2014)	Yes	Visual Arts portfolio	75	14	?	?	?	13	0.95
Newhouse (2014)	Yes	Design portfolio	82	9	?	?	?	13	0.95
Jones, Swan & Pollitt (2015)	No	Maths GCSE scripts	18	12,11	151,150	100%	N/A	~16.7	0.80 0.93
Jones, Swan & Pollitt (2015)	No	Maths task	18	12,11	173,177	114%	N/A	~19.5	0.85 0.93
McMahon & Jones (2014)	No	Chemistry task	154	5	1550	13.2%		~20	0.87

Figure 2.1: Design features* and SSR reliability results from some published CJ/ACJ studies [22]

work [23]. Additionally, frustration also gets created by the lack of students being able to challenge the final results [23].

Many studies' motivation for using adaptivity in CJ studies is to avoid wasting time and resources getting judges to make comparisons whose outcome is a foregone conclusion. However, theoretical considerations from the IRT and CAT literature and the simulation study results in this report show that adaptivity produces spurious scale separation reliability, as indicated by values of the SSR coefficient that are considerably biased upwards from their true value. The higher the proportion of adaptive rounds, the greater the bias. SSR values above 0.70 and even as high as 0.89 can get obtained from random judgments [22].

Consequently, the conclusion is that the SSR statistic is misleading and worthless as an indicator of scale reliability. Other reliability indicators, such as correlations with measures obtained from comparisons among a different group of judges, or correlations with relevant external variables, should be used instead. Therefore ACJ studies that have used high values of the SSR coefficient alone to justify claims that ACJ is a more reliable system than conventional marking need to be re-evaluated [22].

2.2.4 How effective is Comparative Judgement at Providing Feedback?

Multiple studies have got conducted where ACJ has been used to present feedback to the students. The approach gives students insights into how other people have approached a similar situation differently and how peers valued their work [24].

ACJ offers a new way to involve all teachers in summative as well as formative assessment. The model provides strong statistical control to ensure quality assessment for individual students [21].

Synthesising the creation of evidence with its appraisal engaged students in a double looped system of reflection in action. The increase in performance across assignments indicates that students were receiving feedback to support them in improving their work. As this only came from the ACJ judging process, this suggests that students were critiquing their own work relative to the breadth of work presented by their peers and they were also engaged in a critique of the purpose of the design assignments with respect to core competency development. In essence, students were developing, responding to, and applying criteria [25].

However, all these examples allow students to gain feedback in ranked method, of how well they have scored against others with the addition of seeing other students work modelled to them. But at no point are the students getting any truly personalised feedback on what has worked well and what needs improving. Additionally, it relies heavily on students to self-assess and provide their internal improvements, relying on them genuinely understanding the requirements, which would be a meagre chance for less confident, low-achieving students. Therefore, it is a more superficial process and lacks any try impact for methods required in a secondary or sixth form classroom. So we believe that the CJ, while it does remove cognitive loads, actually adds more work for the teacher to provide the basic required information they would need in their classroom to present to the students.

2.3 Other Rating Systems

While comparative judgment has proven to be a suitable method of ranking pairwise matches of students work over the years, it has its limitations. For example, comparative judgment requires every combination to be compared against, which means for a class of thirty students, accounting for four hundred and thirty-five different combinations. Take into account a subject like English, which every student will have to take. A typical school year could have one hundred and twenty students, which would mean seven thousand one hundred and forty different combinations. That is a lot of time and comparisons that would be required. Therefore, to truly take the cognitive load off a marker or teacher, it would be better to try and have different people sub-sample the work. Then, from the scoring of the sub-samples, use this to generate an overall ranking. In essence, it is

creating a competitive scoring system against each other. Two suitable systems to achieve this would be an Elo or Glicko rating system.

2.3.1 Elo Ranking System

The Elo ranking got first introduced into competitive chess in the 1980s [26]. However, it got created in the 1960s by Arpad Elo as a replacement for the Harkness System. The Harkness System got used by the United States Chess Federation (USCF) at that time [27]. Additionally, the Elo system has gets used as a ranking system for football, American football, basketball as well as eSports like Counter-Strike: Global Offensive and League of Legends [28, 29].

The Elo system looks at the difference in two players ratings, then serves as a predictor for the match's outcome. The players Elo rating is depicted as a number and will change over time depending on the games' outcomes, with the winners taking points from the losers. However, how many points get awarded is decided upon the difference in ranking between the players. If the higher ranked player wins, only a few rating points get taken from the lower rank player. However, if an 'upset win' occurs, when the considerably lower rank player beats the higher rank player, then a much greater number of points will be gained to the winner and deducted from the loser. Ultimately, even when 'upset wins' happen, the ranking of the players will reflect the valid scores over time. [find reference]

However, there are ways that players who know how the system works can cheat it. These methods include protecting one's rating, selective pairing and ratings inflation and deflation.

Players protecting one's rating discourages game activity for players wanting to preserve their score. In essence, this situation gets created when players are not playing any more games once they are at a high score **[23]**. A method against this behaviour is to award an activity bonus that gets combined with the ranking score **[24]**.

Selective pairing is when players choose their opponents, which results in players choosing opponents that the player has the minimal risk of losing. Additions like a k-factor got added, but these do not solve the problem completely. **[find reference]**. Additional implementations have been added, like auto-pairing, which are based on random pairings but have a winner stays on context. **[find reference]**.

Inflation is when a score means less over time. For example, a player has a score of 2500 and gets ranked 5, but later, another is ranked 15. It is showing that the player's

ability is decreasing over time. When deflation happens, this indicates that advancement is happening. Deflation is when a score of 2500, got a player ranking of 7, but at a later date, the score is then put ranked the player 2. Therefore, we must consider when using ratings to compare players between different eras. The ranking gets made more difficult when inflation or deflation are present.

The Elo system has a flaw in that it is almost certainly not distributed as a normal distribution. As a result, weaker players have greater winning chances than Elo's model predicts.**[citation needed]** However, the Elo ratings still provides a valuable mechanism for giving a rating based on the opponent's rating.

2.3.2 Glicko Ranking System

The Glicko rating system and Glicko-2 rating system are methods for assessing a player's strength in games of skill, such as chess and Go. It was invented by Mark Glickman as an improvement on the Elo rating system, and initially intended for the primary use as a chess rating system. Glickman's principal contribution to measurement is "ratings reliability", called RD, for ratings deviation.

Both Glicko and Glicko-2 rating systems are under public domain and found implemented on game servers online (like Pokémon Showdown, Lichess, Free Internet Chess Server, Chess.com, Online Go Server (OGS),[1] Counter Strike: Global Offensive, Team Fortress 2,[2] Dota Underlords, Guild Wars 2,[3] Splatoon 2,[4] Dominion Online and Gods Unchained,[5]), and competitive programming competitions. The formulas used for the systems can be found on the Glicko website. The RD measures the accuracy of a player's rating, with one RD being equal to one standard deviation. For example, a player with a rating of 1500 and an RD of 50 has a real strength between 1400 and 1600 (two standard deviations from 1500) with 95% confidence. Twice (exact: 1.96) the RD is added and subtracted from their rating to calculate this range. After a game, the amount the rating changes depends on the RD: the change is smaller when the player's RD is low (since their rating is already considered accurate), and also when their opponent's RD is high (since the opponent's true rating is not well known, so little information is being gained). The RD itself decreases after playing a game, but it will increase slowly over time of inactivity.

2.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of AI that aims to understand natural language through trying to process and analyse it [30, 31]. Ultimately NLP is teaching computers how to understand humans in natural language. However, this is not straightforward, as language is a complex, ever-changing form even for humans. There are three main categories that NLP problems fall into, heuristics, machine learning, and deep learning [31]. The nature of ML algorithms gets designed to work with unknown datasets, allowing data scientists to learn how to use language [30]. While this will bring us a vast amount of insights, as mentioned before, the ever caning landscape of language does not mean that it is perfect and, once made, doesn't need revision. Therefore generating and understanding natural language are the most promising but most challenging tasks in NLP [30, 31].

To understand the complexities of machines attempting to understand language, we must first know what we mean when we state 'what is language'. Language is a structured communication system, which involves many combinations of its fundamental components of varying complexities. For example, some of these components are characters, words and sentences to name a few [31].

Human language gets constructed of four major building blocks, and are phonemes, morphemes, lexemes and syntax, and context [31]. To make an effective NLP app, we need to ensure our application has these different building blocks used within its foundations (see fig: 2.2). However, knowing these building blocks does not entail we can do what we like within NLP. NLP has a lot of challenges that involve ambiguity, common knowledge, creativity and diversity across languages [31].

2.5 Related Work

While comparative judgment is not a new concept, only a few current systems implement a version of it as a tool for marking. These current CJ projects have a slightly different take on the CJ process but have very similar fundamentals. The current offerings are created or provided by RM Compare, a consortium of universities called D-PAC, No More Marking and e-scape.

RM Compare is probably the version with the most prominent presence.

RM Compare uses Adaptive Comparative Judgement (ACJ), based on The Law of Comparative Judgement. The assessor (a teacher, lecturer, examiner or student) is

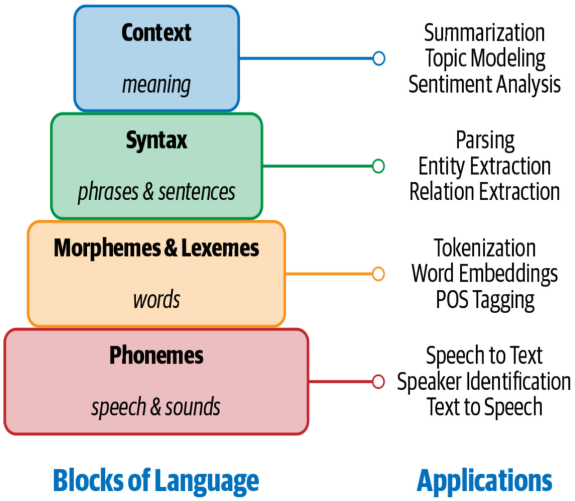


Figure 2.2: The building blocks and their tools for understanding language [31].

presented with two anonymised pieces of work in a side-by-side pairwise comparison and asked to use their professional judgement to select which of the two is better at meeting the assessment criteria.

RM Compare says that through repeated pairwise comparisons, optimised by an iterative, adaptive algorithm, a highly reliable scale or rank order is created through consensus over what 'good', 'better', and 'best' looks.

RM Compare empowers users across educational organisations to collaborate on assessments and is proven to increase student attainment. It also reduces the cognitive load from teachers, which gets achieved through the very nature of the comparative judgment process. It also has a straightforward and effective UI for the user to interact with.

However, it still has an extensive workload as for it to be effective, the markers (known as judges) need to go through several rounds. Multiple examples online were stating 16 rounds. RM Compare states that these numerous rounds are required to reduce the error uncertainty rate. The algorithm's adaptiveness will ensure that pairs closely matched to each other get checked more to confirm the order is correct, reducing the algorithm's error rate calculation. A high level of uncertainty will get compared more often to check the consensus between the judges.

An issue with the application is that it doesn't provide any real form of meaningful feedback. RM Compare suggests that the students gain feedback from the system is for the students to compare their peer's work through the system. Once this comparison by

the students gets completed, the students' peered work ranking results will get compared against the teachers. Which then, in turn, gets used as a point of discussion. Therefore, in our opinion, not providing any meaningful form of feedback. While RM Compare that the process has a considerable impact on students attainment, this claim feels more like a marketing gimmick. While we agree that this process can generate insights into students' expectations, it does not provide meaningful, personalised feedback. Therefore, not allowing them to know what they need to do to improve.

2.5.1 Subsection all similar work

2.5.2 Comparison of similar work

2.6 Overall Aim

Comparative judgement is a power tool. It can remove a lot of cognitive load from the teacher. It also eliminates the teacher's bias in the marking process, especially when the teacher knows whose student they are marking. Teachers can consider how the student has performed over the year instead of how they did in that final piece of work.

However, the current process of adaptive comparative judgment can reduce the cognitive load with the teacher marking and lessen the potential for bias from the teacher. Current implementations do have their limitations and still create a lengthy process. With some systems still having markers to mark student's work up to, some examples have 16 rounds of marking, which is still very time-consuming. Which, if you want to expand this to a national level, wouldn't be very effective.

Therefore, we want to look into different methods of student work ranking orders that could get used to allowing a crowdsourced way of marking in a comparative judgement style, can be implemented. Suggested alternatives are an Elo system ranking. Additionally, we want to create NLP tools that will allow us to interrogate the data and see if there are any patterns within the data and the end rankings. Allowing us to suggest what aspects of the data makes the content get perceived as good.

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

3.1.1 Web 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

Figure 3.1: A visual representation of the processes pipline.

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 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.

3.1.2 NLP Information Extraction

Information extraction is the process of extracting relevant information from text. Some of this information could be calendar events and names of people, to list a few [31]. We, as humans, do this all the time. We extracted the information from multiple sources, like reading documents or conversations. However, for computers, this is not such a straightforward task. Due to the ambiguous nature of natural language, information can mean multiple things depending on the context in which it is getting used.

Due to its complex nature, information extraction relies on several separate takes, which, when used together, generates information. These steps include keyphrase extraction, named entity recognition, named entity disambiguation, and linking and relationship extraction [31].

Next, we will look into the different building blocks that can extract information from our text to provide feedback to the user. We will look into part of speech tagging, named entity recognition, feature extraction, sentiment analysis, text similarity, utterance pattern matching, text similarity scoring and word sequence pattern recognition.

3.1.2.1 Part of Speech Tagging

Part of Speach (POS) tagging has the hidden Markov model (HMM) underpinning it [31]. The HMM is a statistical model that assumes an underlying, unobservable process

with hidden states [32]. POS tagging ultimate aim is to identify the nouns, verbs, and other key parts of speech [30].

We decided to implement POS tagging on the tweets to see if any insights would help provide any feedback to the user. While it might not give us many insights on its own, it can get used as an additional tool that, when paired with other methods, can help provide some insights. We also felt that when the POS tagging got visualised, this would help create a clear picture of the structure of the tweet.

3.1.2.2 Named Entity Recognition

Named Entity Recognition (NER) is the task of identifying entities in a document for information extraction [31]. Entities usually are made up of names of persons, locations, organisations, money expressions and dates, to list a few [33]. NER is an important step within the pipeline of information extraction [31].

As this is a crucial stage in information extraction, we decided to implement it and use it in its pre-trained form from the libraries offerings. We decided to use this method due to the time restrictions of the project and to see how well it performs and if it can help generate feedback to the user.

3.1.2.3 Feature Extraction

Feature extraction aims to transform tokens into features. An excellent technique to achieve this is a bag of words (BOW). This technique will count the occurrences of a particle token within our text. Therefore, for each token, we will have a feature column. This feature column gets referred to as text vectorisation. However, using a standard BOW will lose the word order, and the counters can not be normalised [33].

In order to preserve some order, we can count the tokens as pairs or triplets, for example. This technique gets also referred to as n-grams. The n refers to the number of tokens to get referenced. Some examples are 1-grams for tokens and 2-grams for token pairs. However, this has its problems as it can create too many features [31]. A solution to this problem is to remove some n-grams from the feature set. This solution can get achieved by using the metric based on the frequency of their occurrence [31].

The n-grams that we would want to remove based on their frequency are high and low-frequency n-grams. High-frequency grams get usually referred to as stop words, and low-frequency grams are rare words or typos [33]. We especially want to remove

the low-frequency n-grams as they can create overfitting. Ultimately, we ideally want the medium frequency words.

A technique we can use to find the medium frequency n-grams is term frequency-Inverse document frequency (TF-IDF). TF-IDF has two main stages, the term frequency (TF) and the inverse document frequency (IDF). The TF ($tf(t,d)$) looks for the frequency of the n-gram (term) t in the document d [34]. While IDF takes the total number of documents in the corpus ($N = |D|$) and the number of documents where the term t appears ($|\{d \in D : t \in d\}|$) [34]. So the IDF gets represented as $idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$ [34]. TF-IDF ($tdidf(t,d,D) = tf(t,d) \cdot idf(t,D)$) achieves a high weight by a high-term frequency, within a given document, and a low document frequency of the term in the whole collection of documents [34].

Through using TF-IDF, we can replace counters within our BOW with the TF-IDF value. We can then normalise the result row-wise by dividing by $L_2 - norm$. Through this method, important features will have a relatively high value. Through this method, we are then able to display the key features within our documents.

3.1.2.4 Sentiment Analysis

Sentiment analysis, which can also be known as opinion mining or emotion AI, uses NLP, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis gets widely applied to materials such as reviews and survey responses, online and social media, and healthcare materials for applications. It aims to find out if a perceived text has got classed as positive or negative and, in some instances, neutral [35, 36].

Through aiming to gain an insight into if a tweet is positive or negative can provide some insights into possible patterns emerging. This feature, we believe, could be a helpful tool in providing feedback to the user, especially if there is a clear pattern in terms of a tweets sentiment and its final ranking.

3.1.3 Text Similarity

Text similarity scoring aims to analyse and measure how close two entities of text are to each other [34]. We can compare two objects. By comparing these objects, it is then possible to predict how similar they are. We can use docs, spans, tokens or Lexeme to

calculate the similarity score [**?**]. To measure the similarity scores between text entities, we can use two main types of methods, term and document similarity [34].

Predicting similarity helps build recommendation systems or flag duplicates. For example, it allows for the system to suggest user content that's similar to what they are currently looking at or label a support ticket as a duplicate if it is very similar to an already existing one [**?**]. Additionally, similarity measures are an excellent way to take the noisy text data and group the text together. It allows us to see what text gets considered similar to each other by using unsupervised clustering techniques [34].

As the dataset we are dealing with are Twitter tweets, we decided to do this through entire document similarity and spans of named entities to see if the results provide us with any insights in terms of providing any feedback to the user.

3.1.3.1 Utterence Pattern Matching

3.1.3.2 Finding Word Sequence Patterns

3.1.3.3 Key Phrases

The Key phrases method aims to take a document object and find the word or phrase with the most information to it. This technique is effective, especially when creating a chatbot. Key phrases allow the computer to determine what the user, who is interacting with the chatbot, is talking about. A single word in the question can sometimes be enough, but we might need to look at phrases. Key phrases work well with dependency parsing [30].

We decided to experiment with this feature to see if we could extract the key phrases from the tweets and see if they could provide us with any insights and present them to the user as feedback.

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" [37]. 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 [38].

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 [38]. 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 [38]. 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 [39], Php [40] and JavaScript [41]. We decided to use the Python language [42]. 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 [43], SpaCy [44], Sci-Kit Learn [45], and TensorFlow [46].

3.2.2 Libraries

While we use the Python programming language to create the web application and the NLP information extraction, we require significantly different libraries to complete each task. We will look into the potential web libraries available to us and the NLP focused libraries. We will then present the libraries that we decided upon for each of the parts.

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 [47].

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 [48].

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 [48].

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 packages. 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 [48].

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

There are several NLP library packages already available within Python, all having pros and cons. The most popular and influential libraries are Natural Language Toolkit (NLTK), Gensim, CoreNLP, spaCy, TextBlob, Pattern and PyNLPi.

NLTK is one of the leading platforms for building Python programs that can work with human language data. It presents a practical introduction to programming for language

processing. NLTK comes with a host of text processing libraries for sentence detection, tokenisation, lemmatisation, stemming, parsing, chunking, and POS tagging.NLTK is one of the leading platforms for building Python programs that can work with human language data. It presents a practical introduction to programming for language processing. NLTK comes with a host of text processing libraries for sentence detection, tokenisation, lemmatisation, stemming, parsing, chunking, and POS tagging.

NLTK provides easy-to-use interfaces to over 50 corpora and lexical resources. The tool has the essential functionalities required for almost all kinds of natural language processing tasks with Python.

Gensim is a Python library designed specifically for "topic modeling, document indexing, and similarity retrieval with large corpora." All algorithms in Gensim are memory-independent, w.r.t., the corpus size, and hence, it can process input larger than RAM. With intuitive interfaces, Gensim allows for efficient multicore implementations of popular algorithms, including online Latent Semantic Analysis (LSA/LSI/SVD), Latent Dirichlet Allocation (LDA), Random Projections (RP), Hierarchical Dirichlet Process (HDP) or word2vec deep learning.

Gensim features extensive documentation and Jupyter Notebook tutorials. It largely depends on NumPy and SciPy for scientific computing. Thus, you must install these two Python packages before installing Gensim.

Stanford CoreNLP comprises of an assortment of human language technology tools. It aims to make the application of linguistic analysis tools to a piece of text easy and efficient. With CoreNLP, you can extract all kinds of text properties (like named-entity recognition, part-of-speech tagging, etc.) in only a few lines of code.

Since CoreNLP is written in Java, it demands that Java be installed on your device. However, it does offer programming interfaces for many popular programming languages, including Python. The tool incorporates numerous Stanford's NLP tools like the parser, sentiment analysis, bootstrapped pattern learning, part-of-speech (POS) tagger, named entity recogniser (NER), and coreference resolution system, to name a few. Furthermore, CoreNLP supports four languages apart from English – Arabic, Chinese, German, French, and Spanish.

spaCy is an open-source NLP library in Python. It is designed explicitly for production usage – it lets you develop applications that process and understand huge volumes of text.

spaCy can preprocess text for Deep Learning. It can be be used to build natural language understanding systems or information extraction systems. spaCy is equipped with pre-trained statistical models and word vectors. It can support tokenisation for over 49 languages. spaCy boasts of state-of-the-art speed, parsing, named entity recognition, convolutional neural network models for tagging, and deep learning integration.

TextBlob is a Python (2 & 3) library designed for processing textual data. It focuses on providing access to common text-processing operations through familiar interfaces. TextBlob objects can be treated as Python strings that are trained in Natural Language Processing.

TextBlob offers a neat API for performing common NLP tasks like part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, language translation, word inflection, parsing, n-grams, and WordNet integration.

Pattern is a text processing, web mining, natural language processing, machine learning, and network analysis tool for Python. It comes with a host of tools for data mining (Google, Twitter, Wikipedia API, a web crawler, and an HTML DOM parser), NLP (part-of-speech taggers, n-gram search, sentiment analysis, WordNet), ML (vector space model, clustering, SVM), and network analysis by graph centrality and visualisation.

Pattern can be a powerful tool both for a scientific and a non-scientific audience. It has a simple and straightforward syntax – the function names and parameters are chosen in a way so that the commands are self-explanatory. While Pattern is a highly valuable learning environment for students, it serves as a rapid development framework for web developers.

Pronounced as 'pineapple,' PyNLPl is a Python library for Natural Language Processing. It contains a collection of custom-made Python modules for Natural Language Processing tasks. One of the most notable features of PyNLPl is that it features an extensive library for working with FoLiA XML (Format for Linguistic Annotation).

PyNLPl is segregated into different modules and packages, each useful for both standard and advanced NLP tasks. While you can use PyNLPl for basic NLP tasks like extraction of n-grams and frequency lists, and to build a simple language model, it also has more complex data types and algorithms for advanced NLP tasks.

Although NLTK, TextBlob was used in some experimenting, we decided to use spaCy as the main NLP library. However, NLTK was used on the side (especially with their stop words). As one of the key things we wanted to do was extract information from the tweets, spaCy allowed us to do this and prepare the data for deep learning. While

we did not need a very deep Recurrent Neural Network (RNN), we did implement one to complete the sentiment analysis on the tweets. We used an RNN with two things in mind, to see how well it could perform on small amounts of text, like a tweet, and with the future thoughts of it being able to handle large amounts of text, like someone's essay in an exam. The RNN got constructed by using TensorFlow.

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 without losing its impact.

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.

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

Figure 3.3: Formula to calculate the new Elo score for a tweet.

3.4 Data Set

There were two datasets used within this study. The primary dataset was the ten tweets gathered from Twitter, with a theme of being a joke based on Brexit. The other dataset was the IMDB sentiment analysis dataset. This dataset got used to train and test our RNN model before using our tweets on it.

3.4.1 Data Capture Method

Twitter's developer API got used to allow for the tweets to get extracted. Additionally, the library [name here] got also used. The tweets were then uploaded to the Firebase database through a Python Notebook for the main web app to access them. Having the tweets in the database also allowed us to be then able to create a notebook to then access the data to then do the NLP investigating within.

3.4.2 Pre-Processing

Regarding the data pre-processing within the web app, the only processing occurred was removing the _b characters and replacing them with
 tags. We did this to allow the tweets to have the same layout as they did within Twitter. We decided that a few tweets, especially the Q+A style ones, lost their impact if they were not displayed correctly. Therefore, doing this allowed us to keep the integrity of the tweet and its comedy delivery.

3.4.3 T-Rating Score

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.

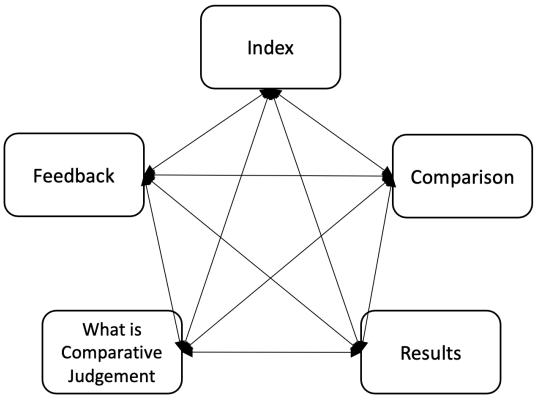


Figure 3.4: A visual representation of the web apps navigation.

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 traditional comparative judgement ranking system. In contrast, the ELO system is using a more traditional approach (see fig: **??**) 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

We will next look at the initial designs compared against the file outcome of the web app. We will also explain the decisions made and what changes we made, and why.

In total, there are five different pages within the web app. The web app has a home page, facilitates the comparative judgement procedure, a results page and a feedback page.

3.6.1 Home Page

3.6.2 Comparison Page

3.6.3 What is Comparative Judgement Page

3.6.4 Results Page

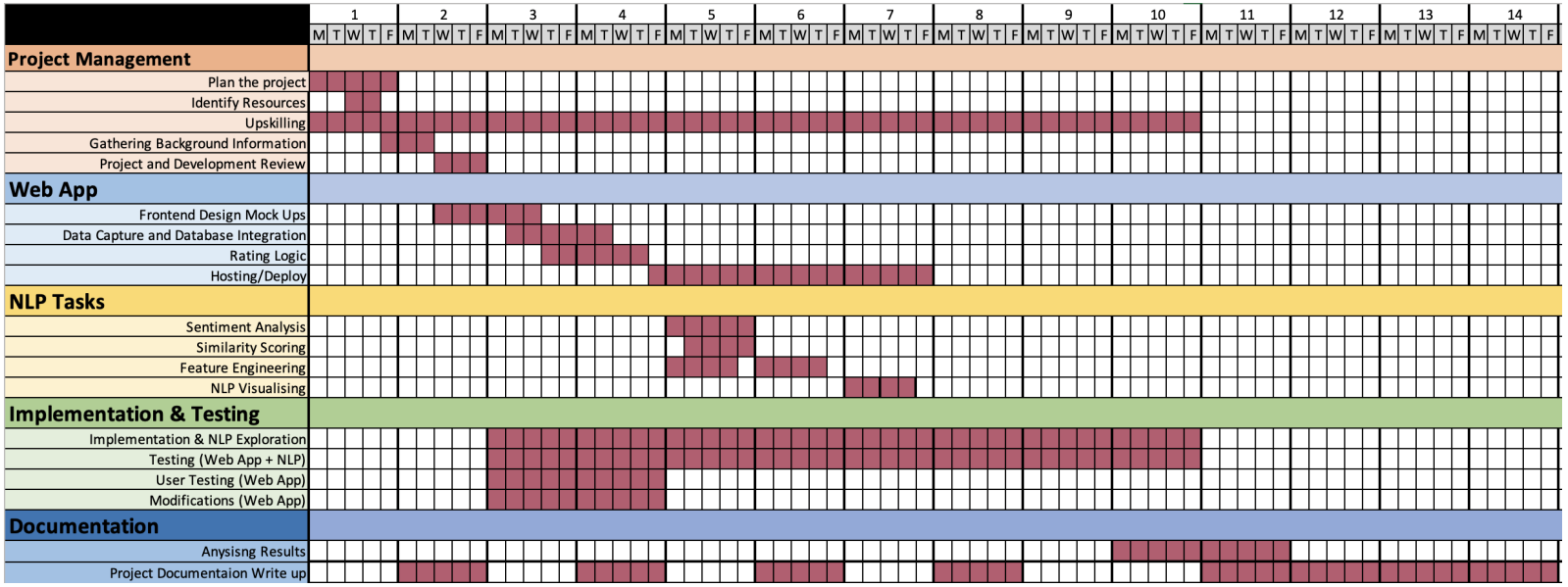
3.6.5 Feedback Page

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


1 2 3 4 5 6 7 8 9 10 11 12 13 14
M T W T F
Project Management
Plan the project
Identify Resources
Upskilling
Gathering Background Information
Project and Development Review
Web App
Frontend Design Mock Ups
Data Capture and Database Integration
Rating Logic
Hosting/Deploy
NLP Tasks
Sentiment Analysis
Similarity Scoring
Feature Engineering
NLP Visualising
Implementation & Testing
Implementation & NLP Exploration
Testing (Web App + NLP)
User Testing (Web App)
Modifications (Web App)
Documentation
Anysisng Results
Project Documentaion Write up

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" [49]. 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 [50]. 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 [50].

The first stage, planning, involves resource allocation, capacity planning, project scheduling, cost estimation, and provisioning [50]. 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 [50]. 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 [50]. 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 [51]. 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 [51].

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

We will first look at the web application results based on the user's feedback, and then we will look into the insights and potential feedback the NLP process could provide the user. We then also look to review the overall process.

We will compare the web application's results against the comparative judgment, Elo ranking, and the score we created for the tweets on Twitter. With the insights of the NLP for feedback to the user, we will look at what insights got made. Additionally, we will look at if any of the knowledge extracted generated provides any meaningful feedback to the user.

4.1 Tweet Ranking Results

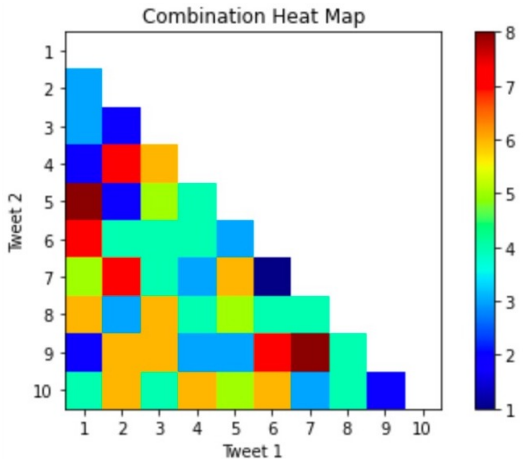


Figure 4.1: The web applicaitons generated results compared agaist each other.

Forty different users take part in the comparison judgement within the web app. Through looking at fig: 4.1 we can see that all combinations got displayed to the users taking part in the comparisons. We can see that tweet one and tweet five appeared the most, while the combination appearing the lowest was tweet six and seven, with one comparison getting presented to the users.

When we look at winners and losers of the comparisons(see fig: 4.2), we can see that the tweet that won the most between a specific combination was tweet four and two, with tweet four winning six times and tweet two winning only once. Additionally, when we look at the combination that appeared the most, one and five, one came out on top five times, compared to five winning between the two once.

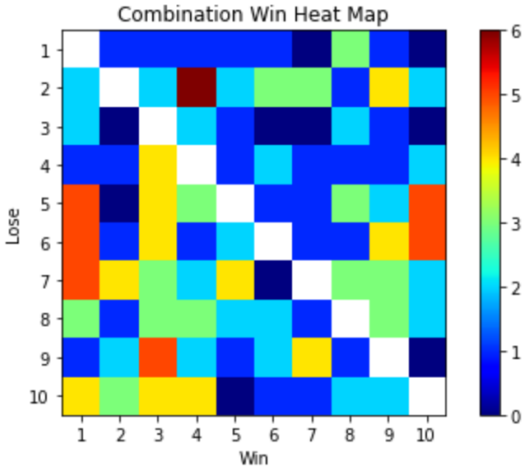


Figure 4.2: A heat map of the amount of times a tweet win or lost

When we look at the winner heat map (see fig: 4.2), we can see that two, five, six, seven and ten had moments where they didn't win a head-to-head with another tweet. Two, six, seven and ten didn't win against at least two different tweets, while the others were only against one tweet they failed to win.

While looking at fig 4.3, we can see that the Elo and comparative judgement ranking generated very similar results. However, as we can see, the tweets coming in 6th, 7th and 8th a slight variation in the results.

While we look at the T-rating ranking compared to the Elo ranking (see fig: 4.4), we can see that the results ranking is very different. The tweet that came first in the T-rankings came fourth in the Elo ranking. At the same time, the tweet that came first in the Elo ranking came eighth in the T-ranking. Tweets that done worse in the Elo ranking compared

Tweet ID	Content	ELO Ranking	ELO Score	CJ Ranking	CJ Score	+/-
3	Q: With Britain leaving the EU how much space was created? A: Exactly 1GB	1	1179.3849804860672	1	21	0
1	An Englishman, a Scotsman and an Irishman walk into a bar. The Englishman wanted to go so they all had to leave. #Brexitjokes	2	1155.592817447446	2	19	0
4	VOTERS: we want to give a boat a ridiculous name UK: no VOTERS: we want to break up the EU and trash the world economy UK: fine	3	1088.8199623047965	3	11	0
9	Hello, I am from Britain, you know, the one that got tricked by a bus	4	997.5535634725744	4	2	0
8	Say goodbye to croissants, people. Delicious croissants. We're stuck with crumpets FOREVER.	5	980.635912446213	6	-2	-1
10	How many Brexiteers does it take to change a light bulb? None, they are all walked out because they didn't like the way the electrician did it.	6	962.7368861475267	5	0	+1
5	#BrexitJokes How did the Brexit chicken cross the road? I never said there was a road. Or a chicken.	7	941.3060728832675	8	-11	-1
2	Why do we need any colour passport? We should just be able to shout, "British! Less of your nonsense!" and stroll straight through.	8	934.560236052883	7	-11	+1
6	After #brexit, when rapper 50 cent performs in GBR he'll appear as 10.000 pounds. #brexitjokes	9	881.9366306271611	9	-14	0
7	I long for the simpler days when #Brexit was just a term for leaving brunch early.	10	877.4729381320648	10	-15	0

Figure 4.3: The web applicaitons generated results compared agaist each other.

to T-ranking had an average difference in the ranked placing of 5 places, while the tweets that had a better Elo ranking compared to the T-ranking ranked an average of 4 places lower. Therefore, 4 of the top 5 tweets in the T-ranking were actually in the bottom five of the Elo ranking. Only tweet ID 4 done one place better with the Elo ranking than it did in the T-ranking. However, two of the top three tweets in the T-ranking were in the bottom three of the Elo ranking and vice versa.

Within the forty participants, twenty-two of them left a justification for why they select one tweet over the other. However, the participants' responses were varied in the amount of provided feedback. Some proved a justification for all five combinations. On the other

Tweet ID	Content	T-Rating Ranking	T-Rating Score	CJ Ranking	+/-
9	Hello, I am from Britain, you know, the one that got tricked by a bus	1	0.57971014	4	-3
2	Why do we need any colour passport? We should just be able to shout, “British! Less of your nonsense!” and stroll straight through.	2	0.20507084	8	-6
6	After #brexit, when rapper 50 cent performs in GBR he'll appear as 10.00 pounds. #brexitjokes	3	0.14233577	9	-6
4	VOTERS: we want to give a boat a ridiculous name UK: no VOTERS: we want to break up the EU and trash the world economy UK: fine	4	0.13602305	3	+1
7	I long for the simpler days when #Brexit was just a term for leaving brunch early.	5	0.05430769	10	-5
8	Say goodbye to croissants, people. Delicious croissants. We're stuck with crumpets FOREVER.	6	0.03097458	5	+1
10	How many Brexiteers does it take to change a light bulb? None, they are all walked out because they didn’t like the way the electrician did it.	7	0.02849923	6	+1
3	Q: With Britain leaving the EU how much space was created? A: Exactly 1GB	8	0.01221757	1	+7
1	An Englishman, a Scotsman and an Irishman walk into a bar. The Englishman wanted to go so they all had to leave. #Brexitjokes	9	0.01165323	2	+7
5	#BrexitJokes How did the Brexit chicken cross the road? "I never said there was a road. Or a chicken".	10	0.00552061	7	+3

Figure 4.4: The Twitter tweet score ranking comparison against Elo ranking.

hand, some only left them for a few and not all. The users gave a total of sixty-three explanations to their decisions on which tweet they had chosen.

One user stated, "I just think it is a clever way to put our departure from EU, plus it did make me giggle." The comment was in regards to tweet three beating tweet eight. Tweet three did provide several justifications, a lot of them to do around its tech theme on Brexit. Some of the rationales are "Comp sci wordplay", "everyone loves a tech joke", "Because it’s the nerdier option", the "First tweet just lol", and "Actually laughed out loud".

Another tweet, tweet ten beating tweet 8, had the justification for winning as 'because of the wordplay'. So we can see that several tweets had some form of explanation around the lines of good wordplay. Therefore, creating user feedback has not made an excellent source of information to help build feedback. However, it has given some context to why they had made their decisions.

4.2 NLP Feedback and Insights

4.3 Overall Results

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.

1: 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] ——, "The method of adaptive comparative judgement," *Assessment in Education: principles, policy & practice*, vol. 19, no. 3, pp. 281–300, 2012.

[22] T. Bramley, "Investigating the reliability of adaptive comparative judgment," *Cambridge Assessment, Cambridge*, vol. 36, 2015.

[23] S. R. Bartholomew, L. Zhang, E. Garcia Bravo, and G. J. Strimel, "A tool for formative assessment and learning in a graphics design course: Adaptive comparative judgement," *The Design Journal*, vol. 22, no. 1, pp. 73–95, 2019.

[24] N. Seery, D. Canty, and P. Phelan, "The validity and value of peer assessment using adaptive comparative judgement in design driven practical education," *International Journal of Technology and Design Education*, vol. 22, no. 2, pp. 205–226, 2012.

[25] N. Seery, J. Buckley, T. Delahunty, and D. Canty, "Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels," *International Journal of Technology and Design Education*, vol. 29, no. 4, pp. 701–715, 2019.

[26] R. C. Weng and C.-J. Lin, "A bayesian approximation method for online ranking." *Journal of Machine Learning Research*, vol. 12, no. 1, 2011.

[27] A. E. Elo, *The Rating of Chessplayers, Past and Present*. New York: Arco Pub., 1978. [Online]. Available: http://www.amazon.com/Rating-Chess-Players-Past-Present/dp/0668047216

[28] N. Silver and R. Fischer-Baum, "How we calculate nba elo ratings," *Dostopno na: http://fivethirtyeight. com/features/how-we-calculate-nba-elo-ratings*, 2015.

[29] S. Pradhan and Y. Abdourazakou, ""power ranking" professional circuit esports teams using multi-criteria decision-making (mcdm)," *Journal of Sports Analytics*, vol. 6, no. 1, pp. 61–73, 2020.

[30] Y. Vasiliev, *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press, 2020.

[31] S. Vajjala, B. Majumder, A. Gupta, and H. Surana, *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media, 2020.

[32] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[33] H. Hapke, C. Howard, and H. Lane, *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Simon and Schuster, 2019.

[34] D. Sarkar, *Text Analytics with python*. Springer, 2016.

[35] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[36] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

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

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

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

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

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

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

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

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

[45] 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.

[46] 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/

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

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

[49] 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.

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

[51] 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.