2019-1002 IST 736 Text Mining

Homework Assignment 5 (week 5)

Ryan Timbrook
NetID: RTIMBROO
Course: IST 736 Text Mining
Term: Fall, 2019
Topic: Training Label Acquisition & Inter-coder Agreement analysis

Table of Contents

1 Introduction

Companies are taking advantage of artificial intelligence (AI) technologies that provide customers with more personalization of their products, which in turn increases engagement and helps to enhance customer loyalty, improving sales. AI is being widely used in industries such as retail, finance, healthcare, automotive, public transportation, and many more. As the possibilities become more clear, and the problem-solving potential increases, this widespread adoption of AI by companies is likely to continue. For AI algorithms to be effective in capturing public sentiment toward a company and their products, annotated training data is required to feed the learning models. A challenge arises in how this data is generated and the accuracy of the labels provided. Recruitment of crowdsourcing, human annotators, through Internet services (e.g., Amazon Mechanical Turk) has become an appealing option that allows multiple labeling tasks to be outsourced in bulk, typically with low overall costs and fast completion rates. With this outsourcing, mechanic comes quality assurance measures that need to be evaluated to ensure the data is being labeled in a uniform, objective manner.

Understanding the public attitudes towards AI and AI governance is needed by any organization wanting to develop new innovative products for their company and customers. How organizations market their AI products and handle their customer data can be positively influenced by understanding their customer base sentiment toward it. To be successful, deployment of AI as of any other new technology very much dependency on the public acceptance of the use of the technology. Without the majority general public use of technology products, data that feeds AI intelligence would be limited and most likely ineffective for its needs. Mining social media channels like Twitter, Facebook, LinkedIn, etc., for public sentiment toward AI can provide companies with the insights necessary to drive their decisions and lead toward more successful product development.

This quote from the 'Center for the Governance of AI, Future of Humanity Institute, Unversity of Oxford' titled: 'Artificial Intelligence: American Attitudes and Trends' provides a concise summary of why mining for the public sentiment on this topic is necessary.

"Advances in artificial intelligence could impact nearly all aspects of society: the labor market, transportation, healthcare, education, and national security. AI's effects may be profoundly positive, but the technology entails risks and disruptions that warrant attention. While technologists and policymakers have begun to discuss AI and applications of machine learning more frequently, public opinion has not shaped much of these conversations. In the U.S., public sentiments have shaped many policy debates, including those about immigration, free trade, international conflicts, and climate change mitigation. As in these other policy domains, we expect the public to become more influential over time. It is thus vital to have a better understanding of how the public thinks about AI and the governance of AI. Such understanding is essential to crafting informed policy and identifying opportunities to educate the public about AI's character, benefits, and risks." - https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/

Being able to effectively messure the quality of crowdsourced sentiment annotations directly impacts the accuracy of the learned sentiment classifier models. One method to ensure this quality is to evaluate how good the turkers (human annotators) are in general. Researchers have done large-scale studies just to evaluate that. A group of authors published a paper in 2008 that demonstrated that if you can hire a large number of non-experts and get their average annotation, that could usually work as well as hiring a small number of experts in those human annotation tasks.

1.1 Purpose

Using the crowdsourcing platform 'Amazon Mechanical Turk' performs an experiment on crowdsourced human sentiment annotation of social media posts to news articles on Artificial Intelligence. Using Cohen's Kappa measurement on the inter-coder agreement, calculate agreement among these multiple coders.

2 Analysis and Models

2.1 About the Data

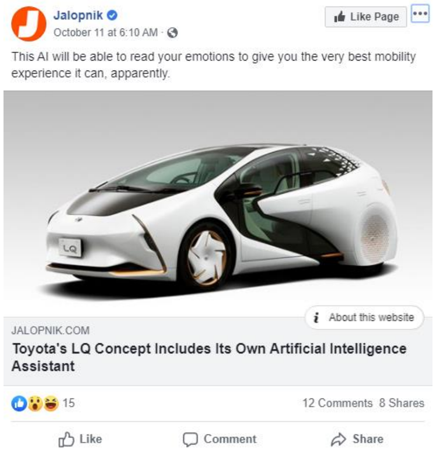
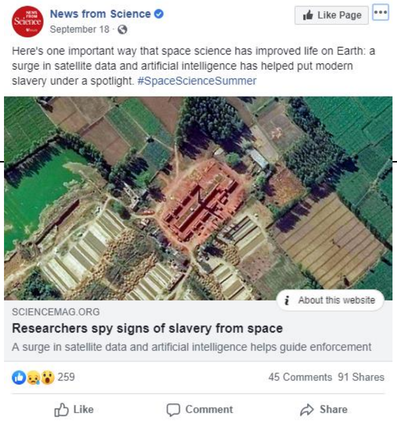
The sentiment data used in this experiment was collected from Facebook public posts on Artificial Intelligence topics from 2019. Four news article headlines were chosen at random to evaluate how crowdsourced human annotators would label public response posts to the news article. For each of the article headlines, three public posts were collected for labeling.

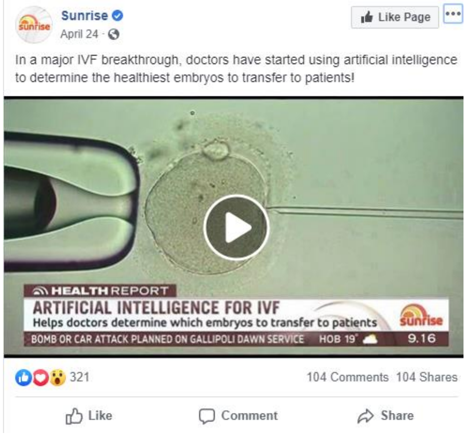
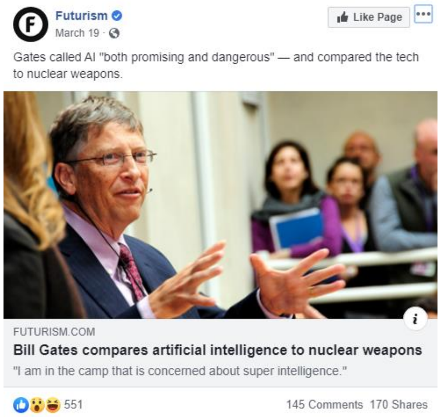
Amazon's Mechanical Turk (MTurk) was used for crowdsourcing labeling of the public Facebook AI posts. MTurk is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development. Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed works over the internet.

2.1.1 *Dataset Aquisition*

2.1.1.1 Facebook News Article Posts

News posts on Artificial Intelligence with a sampling of public response posts.

Facebook Article	Article Headline Text	Public Response Text
 Jalopnik October 11 at 6:10 AM This AI will be able to read your emotions to give you the very best mobility experience it can, apparently. JALOPNIK.COM Toyota's LQ Concept Includes Its Own Artificial Intelligence Assistant 15 12 Comments 8 Shares 	Toyota's LQ Concept Includes Its Own Artificial Intelligence Assistant. This AI will be able to read your emotions to give you the very best mobility experience it can, apparently.	It's like the design team still thinks it's 1992.
		So it disables itself when I want to drive it? Nice
		Yeah it's also wearing a bluetooth ear piece
 News from Science September 18 Here's one important way that space science has improved life on Earth: a surge in satellite data and artificial intelligence has helped put modern slavery under a spotlight. #SpaceScienceSummer SCIENCEMAG.ORG Researchers spy signs of slavery from space A surge in satellite data and artificial intelligence helps guide enforcement 259 45 Comments 91 Shares 	Researchers spy signs of slavery from space. A surge in satellite data and artificial intelligence helps guide	This is my new response to people that tell me we need to "look after people here on Earth before we explore outer space".

		Does it also detect economic slavery from space? That is not a view from space, science these days uses propaganda
 Sunrise ✓ April 24 · Like Page In a major IVF breakthrough, doctors have started using artificial intelligence to determine the healthiest embryos to transfer to patients! HEALTH REPORT ARTIFICIAL INTELLIGENCE FOR IVF Helps doctors determine which embryos to transfer to patients sunrise BOMB OR CAR ATTACK PLANNED ON GALLIPOLI DAWN SERVICE HOB 19° 9.16 321 104 Comments 104 Shares Like Comment Share	In a major IVF breakthrough, doctors have started using artificial intelligence to determine the healthiest embryos to transfer to patients!	I wonder if my cells would have been good enough to be chosen? Wow massive I was lucky to have 2 implanted after many years of ivf and the took resulting in 2 babies who drive me crazy ?? but I love them more every day How unethical!
 Futurism ✓ March 19 · Like Page Gates called AI "both promising and dangerous" — and compared the tech to nuclear weapons. FUTURISM.COM Bill Gates compares artificial intelligence to nuclear weapons "I am in the camp that is concerned about super intelligence." 551 145 Comments 170 Shares	Bill Gates compares artificial intelligence to nuclear weapons. Gates called AI 'both promissing and dangerous' - and compared the tech to nuclear weapons.	Of course it will bite us in the ass eventually, but until then it should be very exciting. pure nonsense! ..it's no different than guns, dependent on how you legislate, it will be safe or end in mass shootings Doesn't sound like something we can ethically handle yet.

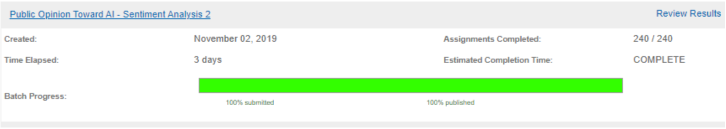
2.1.1.2 Amazon Mechanical Turk

For anyone wanting to use AMT crowdsroucing features they first need to register as an MTurk Request and gain access to the Requester portal. It's also advisable to register as an MTurk Worker to better understand how the workforce will view and consume the requests you make as a Requester. In addition to these production portals, both have Sandbox environment equivalents where you can first create your projects, validating all the content is displayed as expected along with identifying any potential issues with your billing structure prior to releasing it to the workforce. More information on each of these sites can be found below by visiting their links.

- MTurk Requester Site: https://requester.mturk.com/create/projects
- MTurk Request Sandbox Site: https://requestersandbox.mturk.com/create/projects
- MTurk Worker Site: https://workersandbox.mturk.com/
- MTurk Worker Sandbox Site: https://requester.mturk.com/developer/sandbox

The data output for this labeling was produced in two batches. The first batch was the Facebook response posts to the news articles on AI and the second was the news articles headline themselves. The below images are taken from the 'Results' page of the Requester portal.

Figure 2.1: Public Opinion Toward AI - Sentiment Analysis



Public Opinion Toward AI - Sentiment Analysis 2

Review Results

Created: November 02, 2019

Assignments Completed: 240 / 240

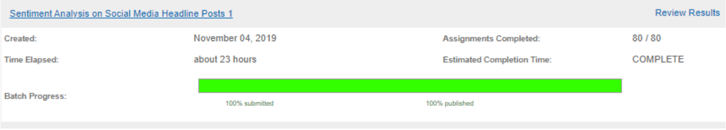
Time Elapsed: 3 days

Estimated Completion Time: COMPLETE

Batch Progress:

100% submitted 100% published

Figure 2.2: Sentiment Analysis on Social Media Headline Posts



Sentiment Analysis on Social Media Headline Posts 1

Review Results

Created: November 04, 2019

Assignments Completed: 80 / 80

Time Elapsed: about 23 hours

Estimated Completion Time: COMPLETE

Batch Progress:

100% submitted 100% published

Each project batch results were downloaded as .csv files for analysis and Kappa evaluation scoring. You access this download option by selecting the 'Review Results' hyperlink show in the above screen images of the 'Results' page.

Project **'Public Opinion Toward AI'** Raw Data Structure:

- Shape: (240, 32)
- Size: 7680

- From this set of data, there are 240 labeled observations. The project was structured such that each of the 12 tasks (i.e., each task represents a single Facebook response post) had a max work assignment of 20 allocated. This allowed for multiple works to be able to label a given task for kappa pair-wise scoring assessment. More details on how this experiment was structured will be discussed in section 2.1.1.3 - Data Experiment Design

Project **'Sentiment Analysis on Social Media Headline Posts'** Raw Data Structure:

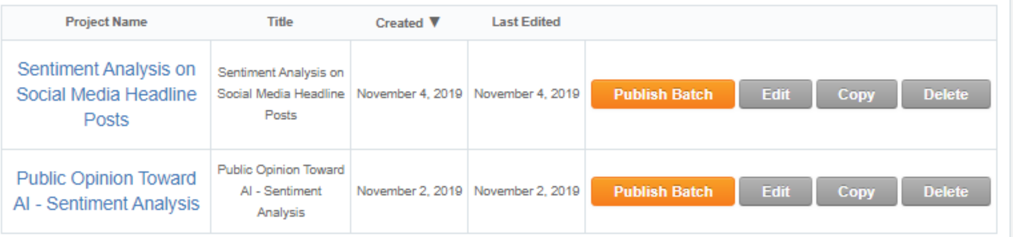
- Shape: (80, 31)
- Size: 2480

- From this set of data, there are 80 labeled observations. The project was structured such that each of the 4 tasks (i.e., each task represents a single Facebook news article headline) had a max work assignment of 20 allocated. This allowed for multiple works to be able to label a given task for kappa pair-wise scoring assessment. More details on how this experiment was structured will be discussed in section 2.1.1.3 - Data Experiment Design

2.1.1.3 Data Experiment Design

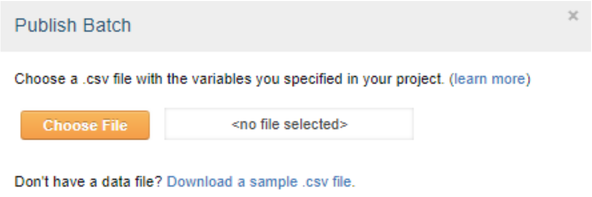
Each of the two experiments was first created and validated in the Requester and Worker Sandbox environments prior to being built and released into the Requester production portal. The text data (Facebook posts) that is defined as the workers' task is uploaded to a created project when you, as the Requester, select the 'Publish Batch' web button shown below in Figure 2.3 and 2.4 below. You can get a sample of the file format by selecting the 'Download a sample .csv file' shown below. During the design layout phase, you can modify the HTML template to include additional variables that would be input through the batch data .csv file. In our first experiment, the Facebook article headline was include with the response post text to give context to the worker when selecting their sentiment label chosen.

Figure 2.3: Mturk Requester - Existing Projects Portal



Project Name	Title	Created ▼	Last Edited	
Sentiment Analysis on Social Media Headline Posts	Sentiment Analysis on Social Media Headline Posts	November 4, 2019	November 4, 2019	Publish Batch · Edit · Copy · Delete
Public Opinion Toward AI - Sentiment Analysis	Public Opinion Toward AI - Sentiment Analysis	November 2, 2019	November 2, 2019	Publish Batch · Edit · Copy · Delete

Figure 2.4: Mturk Requester: Publish Batch



Publish Batch ×

Choose a .csv file with the variables you specified in your project. (learn more)

Choose File   <no file selected>

Don't have a data file? Download a sample .csv file.

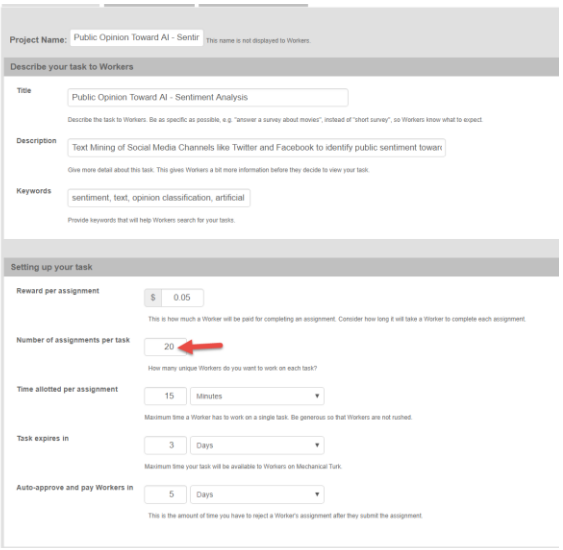
2.1.1.3.1 Public Opinion Toward AI - Sentiment Analysis: Project Design

<u>**Describe task to Workers:**</u>

- **Title:** Public Opinion Toward AI - Sentiment Analysis
- **Description**:
    - Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human-labeled text examples for proper classification of public opinions. Public responses to AI-related posts from both Facebook and Twitter have been collected for sentiment classification tasks.
- **Keywords**: sentiment, text, opinion classification, artificial intelligence, social media posts

<u>**Task Setup:**</u>

- Reward per assignment: **$.05**
- Number of assignments per task: **20**



Project Name: Public Opinion Toward AI - Sentir This name is not displayed to Workers.

Describe your task to Workers

Title: Public Opinion Toward AI - Sentiment Analysis

Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description: Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment towarı

Give more detail about this task. This gives Workers a bit more information before they decide to view your task.

Keywords: sentiment, text, opinion classification, artificial

Provide keywords that will help Workers search for your tasks.

Setting up your task

Reward per assignment: $ 0.05

This is how much a Worker will be paid for completing an assignment. Consider how long it will take a Worker to complete each assignment.

Number of assignments per task: 20

How many unique Workers do you want to work on each task?

Time allotted per assignment: 15 Minutes ▼

Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

Task expires in: 3 Days ▼

Maximum time your task will be available to Workers on Mechanical Turk.

Auto-approve and pay Workers in: 5 Days ▼

This is the amount of time you have to reject a Worker's assignment after they submit the assignment.

- Task Expires in: **3 Days**
- Auto-approve and pay Workers in: **5 Days**

Worker Requirements:

- Require that Workers be Masters to do your tasks: **No**
- Specify any additional qualifications Workers musth meet to work on your tasks: **None**
- The project contains adult content: **No**

Worker Sentiment Analysis Instructions:

These instructions can be modified in the 'Design Layout' HTML form during the project setup steps.

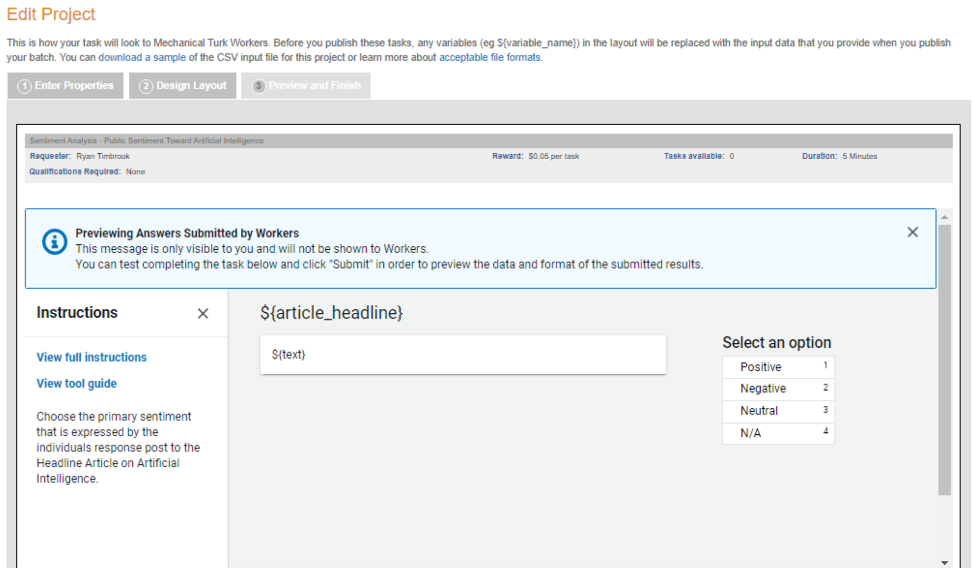
- **Positive** sentiment include: joy, excitement, delight
- **Negative** sentiment include: anger, sarcasm, anxiety
- **Neutral**: neither positive or negative, such as stating a fact
- **N/A**: when the text cannot be understood

When the sentiment is mixed, such as both joy and sadness, use your judgment to choose the stronger emotion.

Design Layout:

Addition instruction were given to the works along with the response posts associated Facebook news article headline to give them context to the statement being made by the Facebook users. The news article headline will appear above the response post text where the ${article_headline} variable is shown below in figure 2.5.

Figure 2.5: Design Preview



Edit Project

This is how your task will look to Mechanical Turk Workers. Before you publish these tasks, any variables (eg ${variable_name}) in the layout will be replaced with the input data that you provide when you publish your batch. You can download a sample of the CSV input file for this project or learn more about acceptable file formats.

Sentiment Analysis - Public Sentiment Toward Artificial Intelligence

Requester: Ryan Timbrook | Reward: $0.05 per task | Tasks available: 0 | Duration: 5 Minutes

Qualifications Required: None

Previewing Answers Submitted by Workers

This message is only visible to you and will not be shown to Workers.

You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Choose the primary sentiment that is expressed by the individuals response post to the Headline Article on Artificial Intelligence.

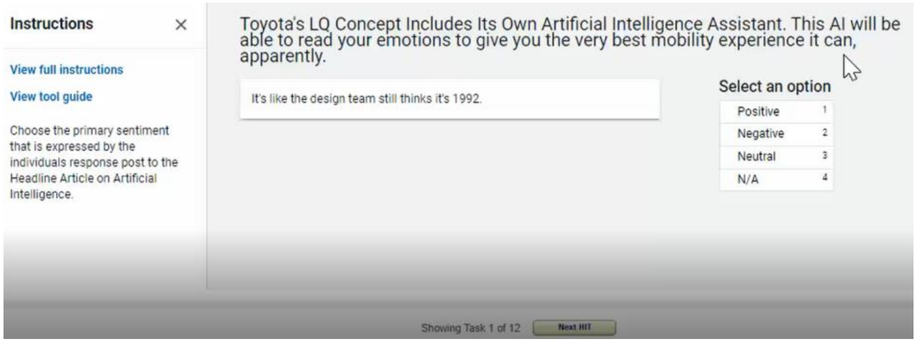
${article_headline}

${text}

Select an option

Additional Instruction text is given to the work shown in the left frame of the design layout.

Figure 2.6: Preview Task Sample



Choose the primary sentiment that is expressed by the individuals response post to the Headline Article on Artificial Intelligence.

Toyota's LQ Concept Includes Its Own Artificial Intelligence Assistant. This AI will be able to read your emotions to give you the very best mobility experience it can, apparently.

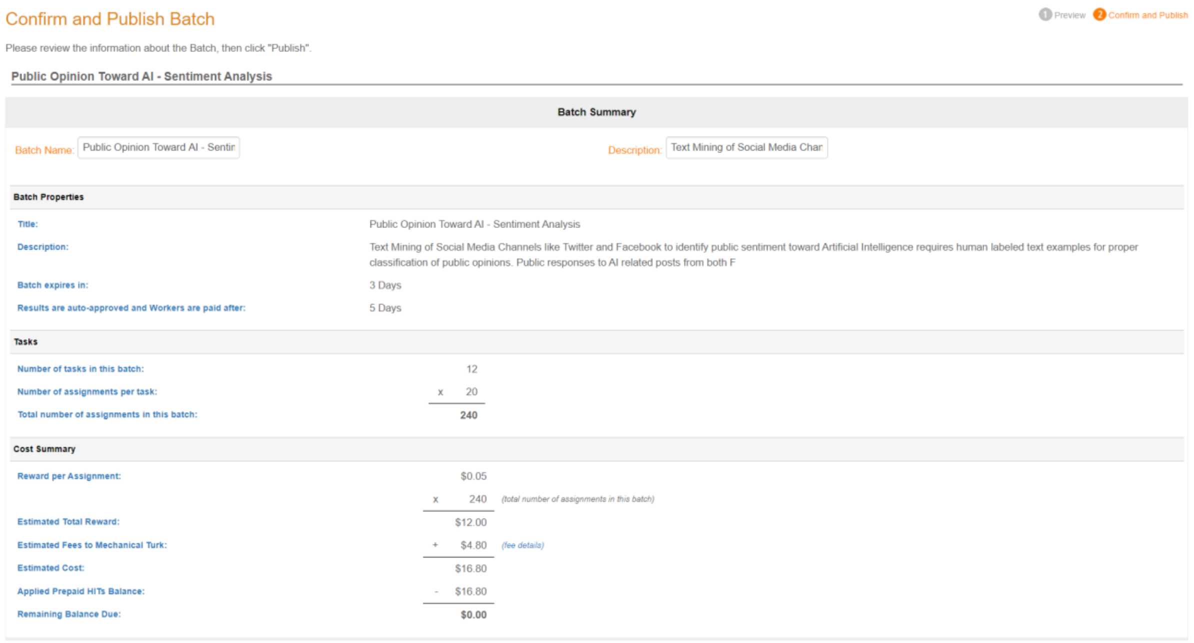
It's like the design team still thinks it's 1992.

Select an option

Showing Task 1 of 12

This view shows how a worker will see an individual task.

2.1.1.3.2 Public Opinion Toward AI - Sentiment Analysis: Project Cost Budget



Confirm and Publish Batch

Please review the information about the Batch, then click "Publish".

Public Opinion Toward AI - Sentiment Analysis

Batch Summary

Batch Name: Public Opinion Toward AI - Sentir

Description: Text Mining of Social Media Char

Batch Properties

Title:	Public Opinion Toward AI - Sentiment Analysis
Description:	Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human labeled text examples for proper classification of public opinions. Public responses to AI related posts from both F
Batch expires in:	3 Days
Results are auto-approved and Workers are paid after:	5 Days

Tasks

Number of tasks in this batch:		12
Number of assignments per task:	x	20
Total number of assignments in this batch:		**240**

Cost Summary

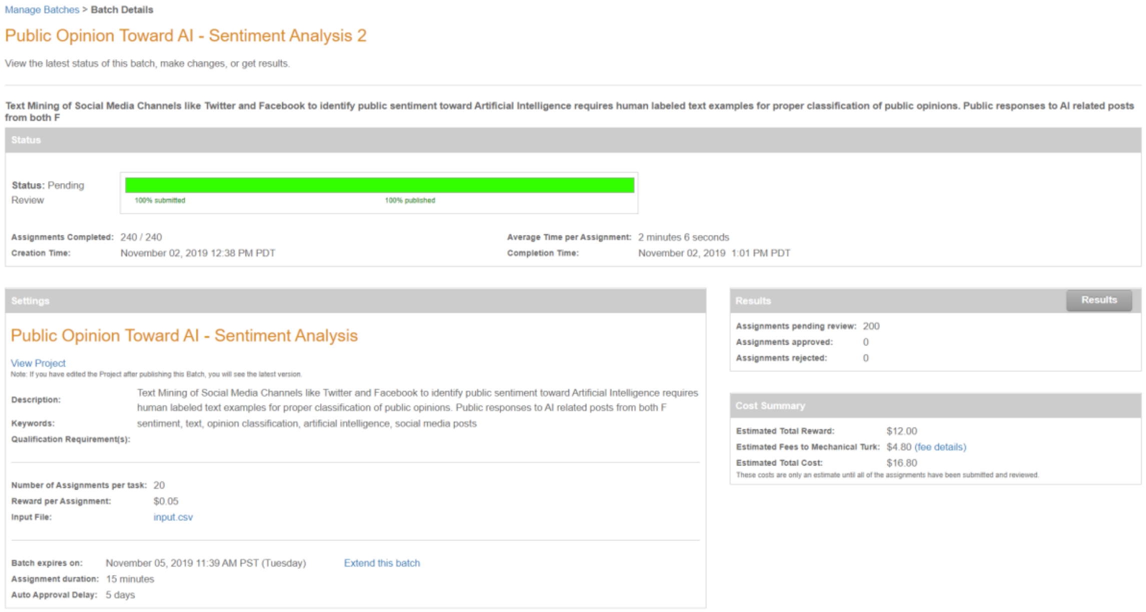
Reward per Assignment:		$0.05	
	x	240	(total number of assignments in this batch)
Estimated Total Reward:		$12.00	
Estimated Fees to Mechanical Turk:	+	$4.80	(fee details)
Estimated Cost:		$16.80	
Applied Prepaid HITs Balance:	-	$16.80	
Remaining Balance Due:		**$0.00**	

2.1.1.3.3 Public Opinion Toward AI - Sentiment Analysis: Project Outcomes

- Project Overall Completion Time: **23 min.**
- Assignments Completed: **240 / 240**
- Average Time per Assignment: **2 minutes 6 seconds**

Outcome Summary:

Figure 2.7: Public Opinion Toward AI Outcome Summary



Manage Batches > **Batch Details**

Public Opinion Toward AI - Sentiment Analysis 2

View the latest status of this batch, make changes, or get results.

Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human labeled text examples for proper classification of public opinions. Public responses to AI related posts from both F

Status

Status: Pending Review

100% submitted 100% published

Assignments Completed:	240 / 240	Average Time per Assignment:	2 minutes 6 seconds
Creation Time:	November 02, 2019 12:38 PM PDT	Completion Time:	November 02, 2019 1:01 PM PDT

Settings

Public Opinion Toward AI - Sentiment Analysis

View Project

Note: If you have edited the Project after publishing this Batch, you will see the latest version.

Description:	Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human labeled text examples for proper classification of public opinions. Public responses to AI related posts from both F
Keywords:	sentiment, text, opinion classification, artificial intelligence, social media posts
Qualification Requirement(s):	

Number of Assignments per task:	20
Reward per Assignment:	$0.05
Input File:	input.csv

Batch expires on:	November 05, 2019 11:39 AM PST (Tuesday)	Extend this batch
Assignment duration:	15 minutes	
Auto Approval Delay:	5 days	

Results

Assignments pending review:	200
Assignments approved:	0
Assignments rejected:	0

Cost Summary

Estimated Total Reward:	$12.00
Estimated Fees to Mechanical Turk:	$4.80 (fee details)
Estimated Total Cost:	$16.80

These costs are only an estimate until all of the assignments have been submitted and reviewed.

Homework Assignment 5 (week 5)

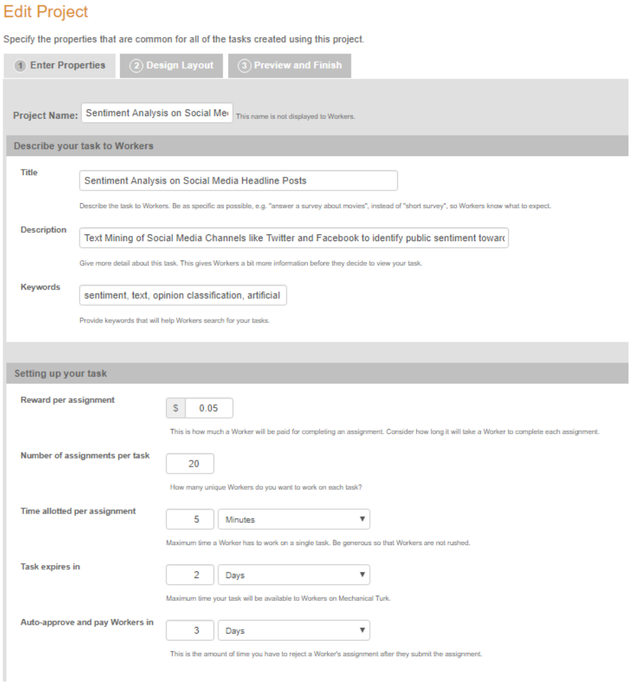
2.1.1.3.4 Sentiment Analysis on Social Media Headline Posts: Project Design

Describe task to Workers:

- **Title:** Sentiment Analysis on Social Media Headline Posts
- **Description**:
  - Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human-labeled text examples for proper classification of public opinions. This grouping of text examples is all news headline posts on Artificial Intelligence. Choose the primary sentiment that is expressed by the news headline post on Artificial Intelligence.
- **Keywords**: sentiment, text, opinion classification, artificial intelligence, social media posts

Task Setup:

- Reward per assignment: **$.05**
- Number of assignments per task: **20**
- Task Expires in: **2 Days**
- Auto-approve and pay Workers in: **3 Days**



Edit Project

Specify the properties that are common for all of the tasks created using this project.

Project Name: Sentiment Analysis on Social Me This name is not displayed to Workers.

Describe your task to Workers

Title: Sentiment Analysis on Social Media Headline Posts

Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description: Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment towar

Give more detail about this task. This gives Workers a bit more information before they decide to view your task.

Keywords: sentiment, text, opinion classification, artificial

Provide keywords that will help Workers search for your tasks.

Setting up your task

Reward per assignment: $ 0.05

This is how much a Worker will be paid for completing an assignment. Consider how long it will take a Worker to complete each assignment.

Number of assignments per task: 20

How many unique Workers do you want to work on each task?

Time allotted per assignment: 5 Minutes

Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

Task expires in: 2 Days

Maximum time your task will be available to Workers on Mechanical Turk.

Auto-approve and pay Workers in: 3 Days

This is the amount of time you have to reject a Worker's assignment after they submit the assignment.

Worker Requirements:

- Require that Workers be Masters to do your tasks: **No**
- Specify any additional qualifications Workers musth meet to work on your tasks: **None**
- Project contains adult content: **No**

Worker Sentiment Analysis Instructions:

These instructions can be modified in the 'Design Layout' HTML form during the project setup steps.

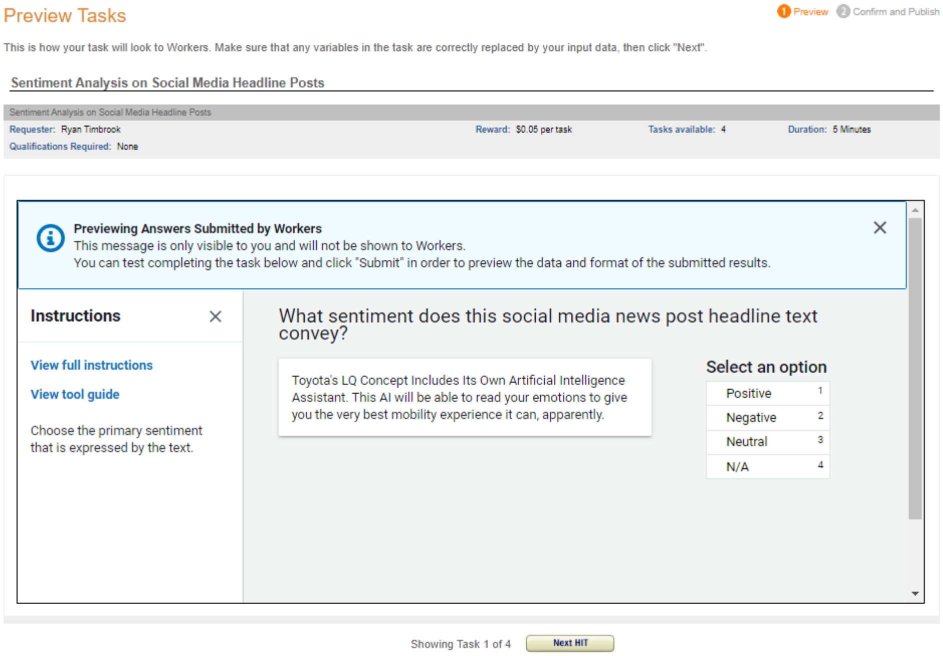
- **Positive** sentiment include: joy, excitement, delight
- **Negative** sentiment include: anger, sarcasm, anxiety
- **Neutral**: neither positive or negative, such as stating a fact
- **N/A**: when the text cannot be understood

When the sentiment is mixed, such as both joy and sadness, use your judgment to choose the stronger emotion.

Homework Assignment 5 (week 5)

Design Layout:

Figure 2.7: Sample Task Worker View



Preview Tasks

This is how your task will look to Workers. Make sure that any variables in the task are correctly replaced by your input data, then click "Next".

Sentiment Analysis on Social Media Headline Posts

Sentiment Analysis on Social Media Headline Posts

Requester: Ryan Timbrook Reward: $0.05 per task Tasks available: 4 Duration: 5 Minutes

Qualifications Required: None

Previewing Answers Submitted by Workers
This message is only visible to you and will not be shown to Workers.
You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Instructions

View full instructions

View tool guide

Choose the primary sentiment that is expressed by the text.

What sentiment does this social media news post headline text convey?

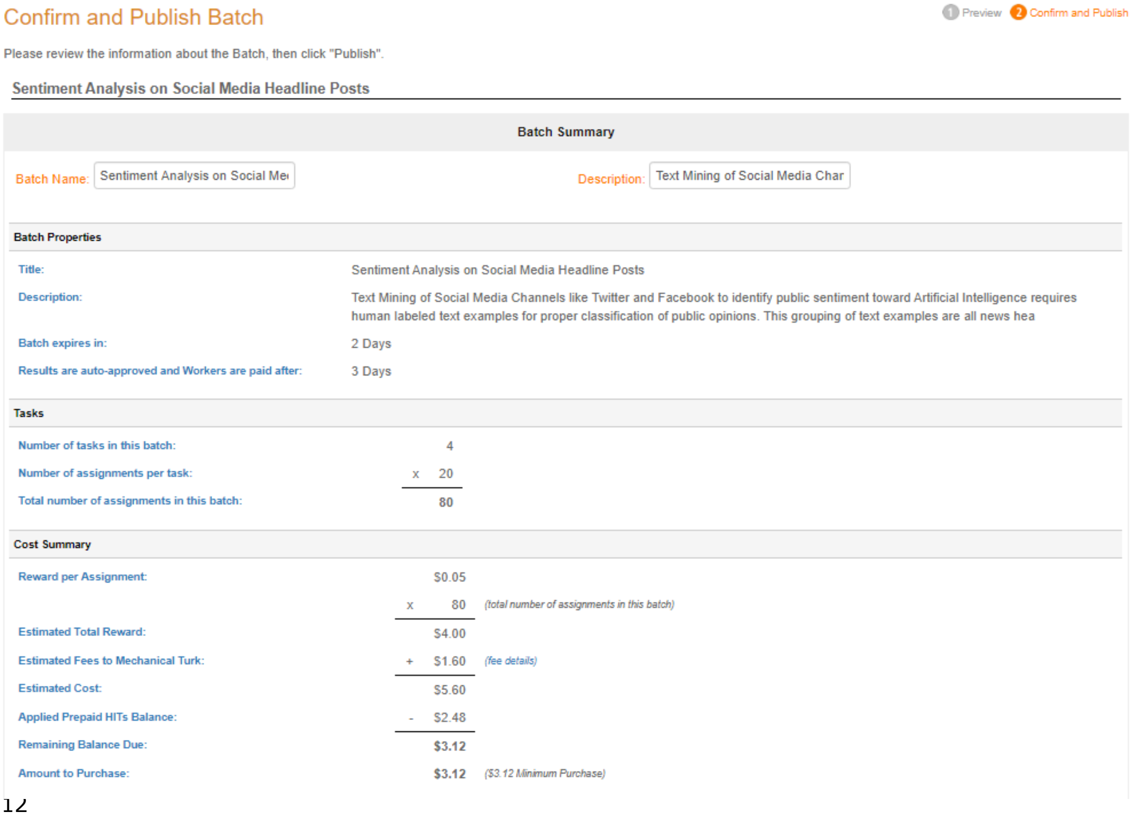
Toyota's LQ Concept Includes Its Own Artificial Intelligence Assistant. This AI will be able to read your emotions to give you the very best mobility experience it can, apparently.

Select an option

Positive	1
Negative	2
Neutral	3
N/A	4

Showing Task 1 of 4

2.1.1.3.5 Sentiment Analysis on Social Media Headline Posts: Project Cost Budget



Confirm and Publish Batch

Please review the information about the Batch, then click "Publish".

Sentiment Analysis on Social Media Headline Posts

Batch Summary

Batch Name: Sentiment Analysis on Social Me

Description: Text Mining of Social Media Char

Batch Properties

Title:	Sentiment Analysis on Social Media Headline Posts
Description:	Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human labeled text examples for proper classification of public opinions. This grouping of text examples are all news hea
Batch expires in:	2 Days
Results are auto-approved and Workers are paid after:	3 Days

Tasks

Number of tasks in this batch:		4
Number of assignments per task:	x	20
Total number of assignments in this batch:		80

Cost Summary

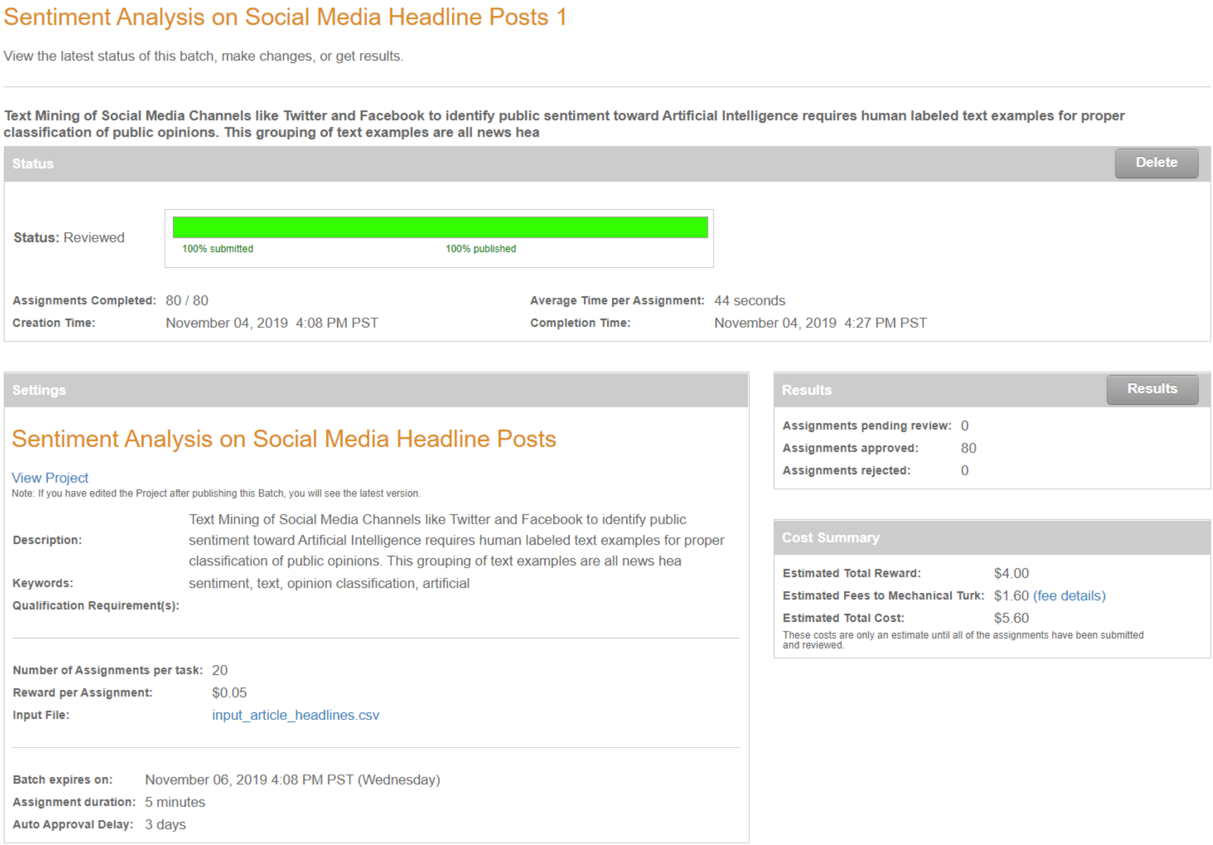
Reward per Assignment:		$0.05	
	x	80	(total number of assignments in this batch)
Estimated Total Reward:		$4.00	
Estimated Fees to Mechanical Turk:	+	$1.60	(fee details)
Estimated Cost:		$5.60	
Applied Prepaid HITs Balance:	-	$2.48	
Remaining Balance Due:		**$3.12**	
Amount to Purchase:		**$3.12**	($3.12 Minimum Purchase)

2.1.1.3.6 Sentiment Analysis on Social Media Headline Posts: Project Outcomes

- Project Overall Completion Time: **19 min.**
- Assignments Completed: **80 / 80**
- Average Time per Assignment: **44 seconds**

<u>Outcome Summary:</u>

Figure 2.8: Sentiment Analysis on Social Media Headline Posts Outcome Summary



Sentiment Analysis on Social Media Headline Posts 1

View the latest status of this batch, make changes, or get results.

Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human labeled text examples for proper classification of public opinions. This grouping of text examples are all news hea

Status: Reviewed

Assignments Completed:	80 / 80	**Average Time per Assignment:**	44 seconds
Creation Time:	November 04, 2019 4:08 PM PST	**Completion Time:**	November 04, 2019 4:27 PM PST

Sentiment Analysis on Social Media Headline Posts

View Project

Note: If you have edited the Project after publishing this Batch, you will see the latest version.

Description:	Text Mining of Social Media Channels like Twitter and Facebook to identify public sentiment toward Artificial Intelligence requires human labeled text examples for proper classification of public opinions. This grouping of text examples are all news hea
Keywords:	sentiment, text, opinion classification, artificial
Qualification Requirement(s):	

Number of Assignments per task:	20
Reward per Assignment:	$0.05
Input File:	input_article_headlines.csv

Batch expires on:	November 06, 2019 4:08 PM PST (Wednesday)
Assignment duration:	5 minutes
Auto Approval Delay:	3 days

Assignments pending review:	0
Assignments approved:	80
Assignments rejected:	0

Estimated Total Reward:	$4.00
Estimated Fees to Mechanical Turk:	$1.60 (fee details)
Estimated Total Cost:	$5.60

These costs are only an estimate until all of the assignments have been submitted and reviewed.

3 Models

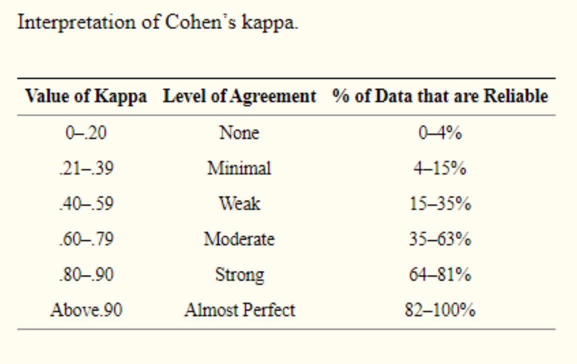
3.1.1 *Cohen's Kappa Model*

Chohen's kappa is a statistic measure of inter-annotator agreeement. To evaluate the level of agreement between the Amazon Mturk workers annotating our sentiment datasets we use the sklearn.metrics.cohen_kappa_score class to perform the calculations. This function computes Cohen's kappa, a score that expresses the level of agreement between two annotators on a classification problem.

Kappa is the proportion of agreement corrected for chance, and scaled to vary from -1 to+ 1 so that a negative value indicates poorer than chance agreement, zero indicates exactly chance agreement, and a positive value indicates better than chance agreement. A value of unity indiecates perfect agreement. The use of kappa implicitly assumes that all disagreemts are equally serious.

Because we have multiple workers coding the same documents we calculate all coder pairs then take the overall average, giving the average pairwise kappa score for the dataset. Below are two figures taken from a Biochemida medica article on kappa statistic. - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Figure 3.1: Interpreting Cohen's Kappa



Interpretation of Cohen's kappa.

Value of Kappa	Level of Agreement	% of Data that are Reliable
0–.20	None	0–4%
.21–.39	Minimal	4–15%
.40–.59	Weak	15–35%
.60–.79	Moderate	35–63%
.80–.90	Strong	64–81%
Above.90	Almost Perfect	82–100%

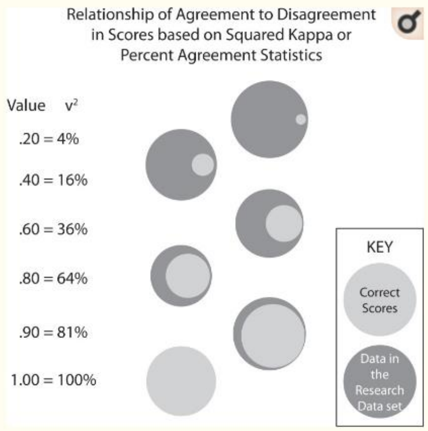


Figure 2.

Graphical representation of amount of correct data by % agreement or squared kappa value.

3.1.1.1 Pair-Wise Kappa Model - Public Opinion Toward AI Response Posts

This model comparies the inter-coder agreement of the twelve Facebook user response posts to AI related topics described in section 2.1.

3.1.1.1.1 Pair-Wise Kappa Details

Of the original 240 dataset samples collected, 120 were kept for this kappa pair-wise calculation. The dataset was narrowed down to keep ony those observations where a unique worker had completed all twelve tasks. This allows us to compute the agreement among raters.

41 unique Mturk workers completed tasks for this assignment, of those 10 Mturk works completed all twelve of the classification tasks. It is these 10 Mturk works the kappa statistic was applied.

Below are two bar graphs that show that after narrowing the observation set to the 10 Mturk workers, the average scoring for each the tasks changed little, if at all.

Table 3.1: Label Mapping

label_id	label_value
0	N/A
1	Negative
2	Neutral
3	Positive

Figure 3.2: All 240 observations over 41 Mturks

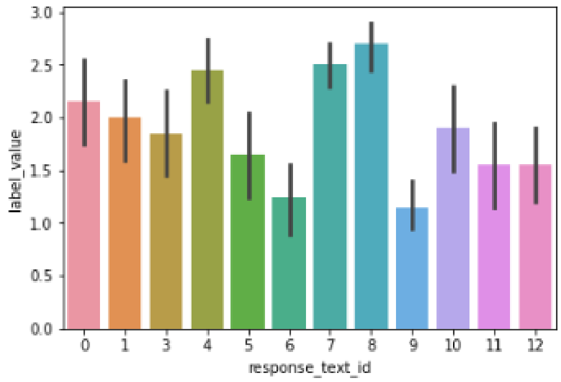


Figure 3.3: Narrowed 120 observations over 10 Mturks

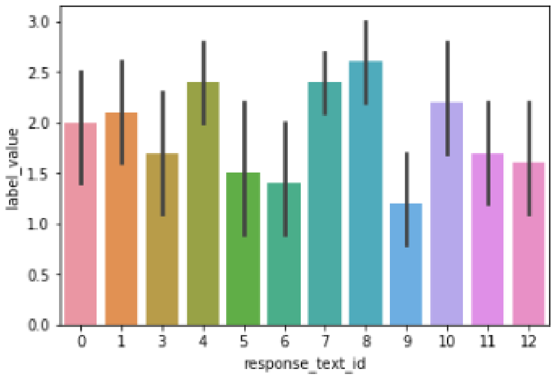


Table 3.2: Response Text Mapping

respons_text_id	text
1	"It's like the design team still thinks it's 1992."
2	"So it disables itself when I want to drive it? Nice"
3	"Yeah it's also wearing a bluetooth ear piece"
4	'This is my new response to people that tell me we need to "look after people here on Earth before we explore outer space".'
5	"Does it also detect economic slavery from space?"
6	"That is not a view from space, science these days uses propaganda"
7	"I wonder if my cells would have been good enough to be chosen?"
8	"Wow massive I was lucky to have 2 implanted after many years of ivf and the took resulting in 2 babies who drive me crazy ?? but I love them more every day"
9	"How unethical!"

10	"Of course it will bite us in the ass eventually, but until then it should be very exciting."
11	"pure nonsense! ..it's no different than guns, dependent on how you legislate, it will be safe or end in mass shootings\nswiss has the safest gun control!!"
12	"Doesn't sound like something we can ethically handle yet."

3.1.1.1.2 Pair-Wise Kappa Results

45 pair-wise agreement calculations were averaged giving the following results:

- Kappa PAIR-WISE AVG SCORE: Public Opinion Toward AI: **0.08170260903575165**
- Kappa PAIR-WISE MEDIAN SCORE: Public Opinion Toward AI: **0.08474576271186429**

According to figure 3.1, Interpreting Cohen Kappa scores, the pair-wise scoring for this data set reflects no agreement.

3.1.1.2 Pair-Wise Kappa Model - Sentiment Analysis on Social Media Headline Posts

This model compares the inter-coder agreement of the four Facebook news article headline posts on AI related topics described in section 2.2.

3.1.1.2.1 Pair-Wise Kappa Details

Of the original 80 dataset samples collected, 53 were kept for this kappa pair-wise calculation. The dataset was narrowed down to keep only those observations where a unique worker had completed all four tasks. This allows us to compute the agreement among raters.

28 unique Mturk workers completed tasks for this assignment, of those 14 Mturk works completed all four of the classification tasks. It is these 14 Mturk works the kappa statistic were applied.

Below are two bar graphs that show that after narrowing the observation set to the 14 Mturk workers, the average scoring for each of the tasks changed little, if at all.

Table 3.1: Label Mapping

label_id	label_value
0	N/A
1	Negative
2	Neutral
3	Positive

Figure 3.2: All 80 observations over 41 Mturks

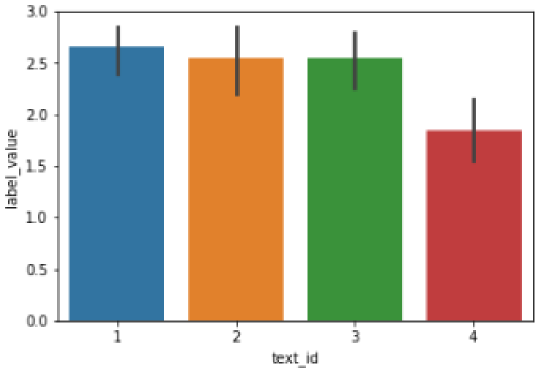


Figure 3.3: Narrowed 53 observations over 14 Mturks

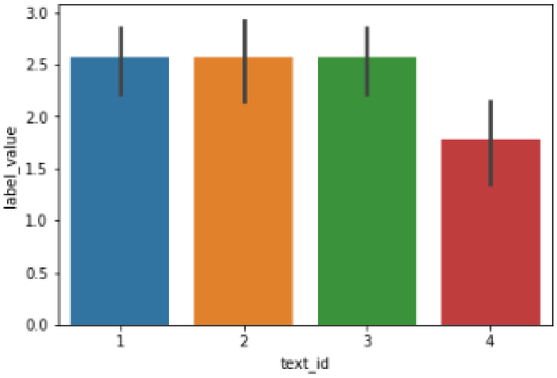


Table 3.2: Text Id Mapping

text_id	text
1	"Toyota's LQ Concept Includes Its Own Artificial Intelligence Assistant. This AI will be able to read your emotions to give you the very best mobility experience it can, apparently."
2	"Researchers spy signs of slavery from space. A surge in satellite data and artificial intelligence helps guide enforcement. Here's one important way that space science has improved life on Earth: a surge in satellite data and artificial intelligence has helped put modern slavery under a spotlight."
3	"In a major IVF breakthrough, doctors have started using artificial intelligence to determine the healthiest embryos to transfer to patients!"
4	"Bill Gates compares artificial intelligence to nuclear weapons. Gates called AI 'both promising and dangerous' - and compared the tech to nuclear weapons."

3.1.1.2.2 Pair-Wise Kappa Results

91 pair-wise agreement calculations were averaged giving the following results:

- Kappa PAIR-WISE AVG SCORE:: Sentiment Analysis on Social Media Headline Posts: **0.07964733679**
- Kappa PAIR-WISE MEDIAN SCORE: Sentiment Analysis on Social Media Headline Posts: **0.11111111**

According to figure 3.1, Interpreting Cohen Kappa scores, the pair-wise scoring for this data set reflects no agreement.

4 Conclusion

Basing our results solely on the pair-wise kappa statistic results, the data collected by crowdsourcing through Amazon Mturk would be considered statistically insignificant. Additional exploratory analysis should be followed up on in these datasets. The raw pair-wise scoring output showed there are a number of kappa scores with large negative values. A negative kappa represents agreement worse than expected, or disagreement. Low negative values (0 to -10) may generally be interpreted as 'no agreement'. A large negative kappa represents great disagreement among raters. Data collected under conditions of such disagreement among raters are not meaningful. They are more like random data than properly collected research data or quality clinical laboratory readings. Those data are unlikely to represent the facts of the situation with any meaningful degree of accuracy. Such a finding requires action to either retrain raters or redesign the instrument.

Compared to automatic sentiment analysis and manual methods, crowdsourcing sentiment analysis of the written text has been shown to achieve high accuracy in many scientific studies. Amazon Mturk is a platform that should be explored further. There are many Mturk worker requirements that could be configured to allow for more accurate results as well as enhancements that can be made to the design layouts of the sentiment analysis questions that would better set the context for the classification tasks.