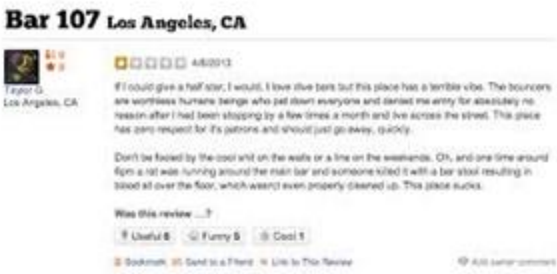
Latif, Aneeka
IST 736
HW 5

Introduction

Living in San Francisco, it's almost as if new restaurants pop up every day. The best way to hear about these restaurants are either through the news or restaurant related apps such as Yelp! that serve as a database and promote restaurants based on user reviews. The average user expects that these reviews to be true renditions of real experiences but unfortunately this is not the case.

Yelp especially has been in hot water over the years due to accusations of untrue positive reviews. Yelp has since taken some action on preventing fake reviews or extremely polarized reviews (Hawkins, 2018[1]) but will not be able to catch all instances of false reviews, for example, in the case of employees boosting their own ratings. This is unfortunate because employees have very clear incentives to want good online postings, they could have promotions depending on their perceived customer service or could be coerced into doing so by managers and employers. Naturally, the employer's incentive in having good reviews, or rather, not having poor reviews is directly correlated with the amount of foot traffic and customers that will enter their establishment; for the employer this is free advertisement. Users are not looking for poor experiences, they rely on word of mouth and word-of-the-internet to discover must-see eateries. Once a user's read a bad review, it's highly unlikely for them to proceed to dine there.



Bar 107 Los Angeles, CA

Taylor G.
Los Angeles, CA

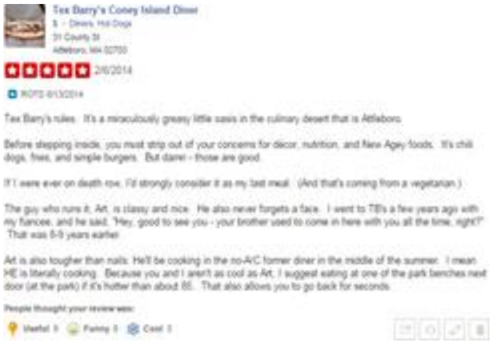
4/6/2013

If I could give a half star, I would. I love dive bars but this place has a terrible vibe. The bouncers are worthless humans beings who pat down everyone and denied me entry for absolutely no reason after I had been stopping by a few times a month and live across the street. This place has zero respect for it's patrons and should just go away, quickly.

Don't be fooled by the cool shit on the walls or a line on the weekends. Oh, and one time around 8pm a rat was running around the main bar and someone killed it with a bar stool resulting in blood all over the floor, which wasn't even properly cleaned up. This place sucks.

Was this review ...?

Useful 8 Funny 5 Cool 1



Tex Barry's Coney Island Diner
$ · Diners, Hot Dogs
31 County St
Attleboro, MA 02703

2/6/2014

ROTD 6/13/2014

Tex Barry's rules. It's a miraculously greasy little oasis in the culinary desert that is Attleboro.

Before stepping inside, you must strip out of your concerns for decor, nutrition, and New Agey foods. It's chili dogs, fries, and simple burgers. But damn - those are good.

If I were ever on death row, I'd strongly consider it as my last meal. (And that's coming from a vegetarian.)

The guy who runs it, Art, is classy and nice. He also never forgets a face. I went to TB's a few years ago with my fiancee, and he said, "Hey, good to see you - your brother used to come in here with you all the time, right?" That was 8-9 years earlier.

Art is also tougher than nails. He'll be cooking in the no-A/C former diner in the middle of the summer. I mean, HE is literally cooking. Because you and I aren't as cool as Art, I suggest eating at one of the park benches next door (at the park) if it's hotter than about 85. That also allows you to go back for seconds.

People thought your review was:

Useful 1 Funny 1 Cool 1

Figure 1: Poor restaurant review for a bar in Los Angeles

Figure 2: Great restaurant review for a diner in Brooklyn

Analysis and Models

About the Data

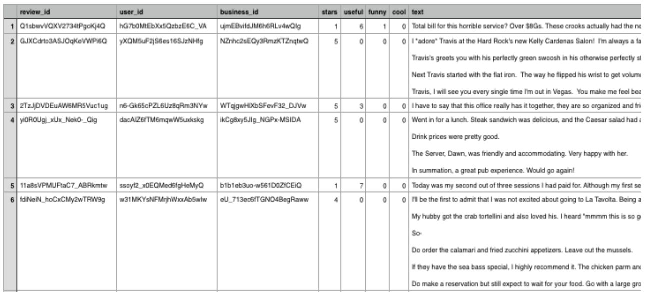
The dataset is a csv file comprised of 25 yelp reviews. Each row in the file consists of the full user review in text format, review_id, user_id, business_id, star_rating, date, and integer values representing how many users marked a particular review as useful, funny, or cool, respectively. The dataset was reduced significantly to expedite annotations from human raters. The text data was isolated for submission to Amazon Mechanical Turk.

Crowdsourcing Annotations

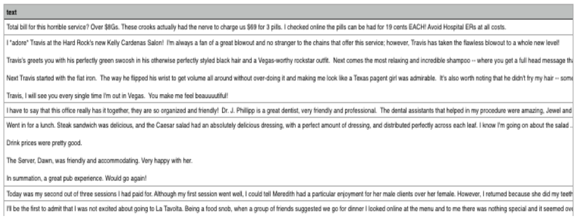
[1] Hawkins, Joy; Yelp vs Google: How they deal with fake reviews

Latif, Aneeka
IST 736
HW 5

Amazon Mechanical Turk is a cost effective and time effective method of crowd sourcing annotations. Rather than annotating a data set yourself, you can submit a dataset to AMT, split up the dataset into smaller sets and pay others to do the labeling on your behalf. The only caveat is making sure the annotators provide quality work and aren't spamming or rushing through the set too quickly.



	review_id	user_id	business_id	stars	useful	funny	cool	text
1	Q1sbwvVQXV2734tPgoKj4Q	hG7b0MtEbXx5QzbzE6C_VA	ujmEBvifdJM6h6RLv4wQIg	1	6	1	0	Total bill for this horrible service? Over $8Gs. These crooks actually had the ne
2	GJXCdrto3ASJOqKeVWPi6Q	yXQM5uF2jS6es16SJzNHfg	NZnhc2sEQy3RmzKTZnqtwQ	5	0	0	0	I *adore* Travis at the Hard Rock's new Kelly Cardenas Salon! I'm always a fa Travis's greets you with his perfectly green swoosh in his otherwise perfectly st Next Travis started with the flat iron. The way he flipped his wrist to get volume Travis, I will see you every single time I'm out in Vegas. You make me feel bea
3	2TzJjDVDEuAW6MR5Vuc1ug	n6-Gk65cPZL6Uz8qRm3NYw	WTqjgwHlXbSFevF32_DJVw	5	3	0	0	I have to say that this office really has it together, they are so organized and fri
4	yi0R0Ugj_xUx_Nek0-_Qig	dacAIZ6fTM6mqwW5uxkskg	ikCg8xy5JIg_NGPx-MSIDA	5	0	0	0	Went in for a lunch. Steak sandwich was delicious, and the Caesar salad had a Drink prices were pretty good. The Server, Dawn, was friendly and accommodating. Very happy with her. In summation, a great pub experience. Would go again!
5	11a8sVPMUFtaC7_ABRkmtw	ssoyf2_x0EQMed6fgHeMyQ	b1b1eb3uo-w561D0ZfCEiQ	1	7	0	0	Today was my second out of three sessions I had paid for. Although my first se
6	fdiNeiN_hoCxCMy2wTRW9g	w31MKYsNFMrjhWxxAb5wIw	eU_713ec6fTGNO4BegRaww	4	0	0	0	I'll be the first to admit that I was not excited about going to La Tavolta. Being a My hubby got the crab tortellini and also loved his. I heard "mmmm this is so go So- Do order the calamari and fried zucchini appetizers. Leave out the mussels. If they have the sea bass special, I highly recommend it. The chicken parm and Do make a reservation but still expect to wait for your food. Go with a large gro



text
Total bill for this horrible service? Over $8Gs. These crooks actually had the nerve to charge us $69 for 3 pills. I checked online the pills can be had for 19 cents EACH! Avoid Hospital ERs at all costs.
I *adore* Travis at the Hard Rock's new Kelly Cardenas Salon! I'm always a fan of a great blowout and no stranger to the chains that offer this service; however, Travis has taken the flawless blowout to a whole new level! Travis's greets you with his perfectly green swoosh in his otherwise perfectly styled black hair and a Vegas-worthy rockstar outfit. Next comes the most relaxing and incredible shampoo -- where you get a full head message th Next Travis started with the flat iron. The way he flipped his wrist to get volume all around without over-doing it and making me look like a Texas pagent girl was admirable. It's also worth noting that he didn't fry my hair -- some Travis, I will see you every single time I'm out in Vegas. You make me feel beauuuutiful!
I have to say that this office really has it together, they are so organized and friendly! Dr. J. Phillipp is a great dentist, very friendly and professional. The dental assistants that helped in my procedure were amazing, Jewel and
Went in for a lunch. Steak sandwich was delicious, and the Caesar salad had an absolutely delicious dressing, with a perfect amount of dressing, and distributed perfectly across each leaf. I know I'm going on about the salad ... Drink prices were pretty good. The Server, Dawn, was friendly and accommodating. Very happy with her. In summation, a great pub experience. Would go again!
Today was my second out of three sessions I had paid for. Although my first session went well, I could tell Meredith had a particular enjoyment for her male clients over her female. However, I returned because she did my teeth
I'll be the first to admit that I was not excited about going to La Tavolta. Being a food snob, when a group of friends suggested we go for dinner I looked online at the menu and to me there was nothing special and it seemed ov

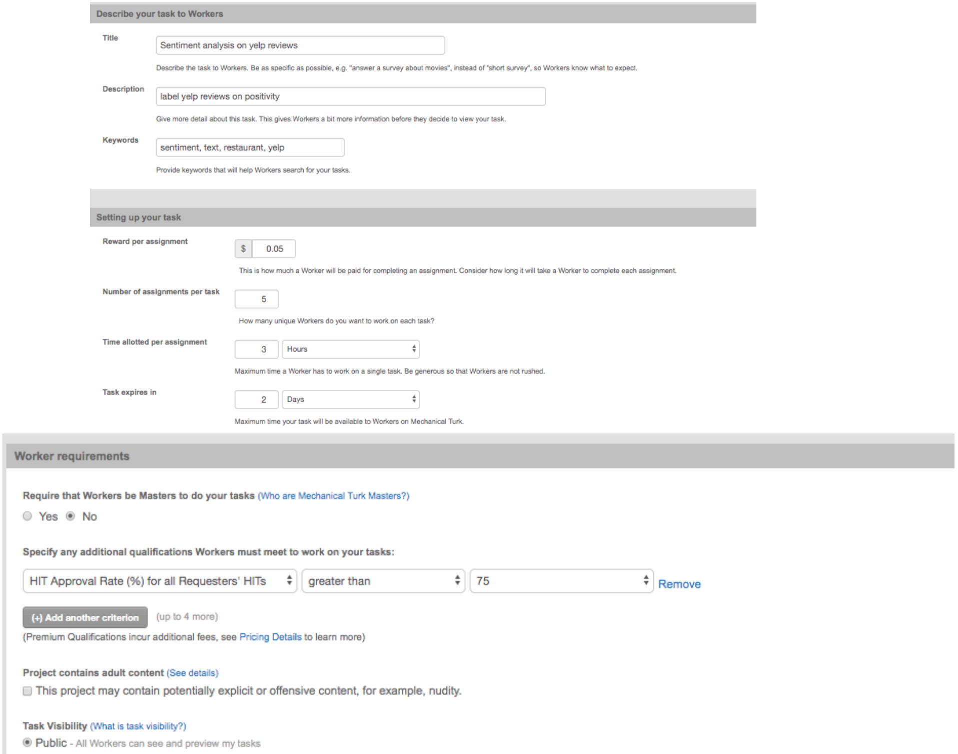
Figure 3: original CSV file

Figure 4: sample set with isolated reviews for AMT

The smaller sample set of 25 reviews was uploaded to Amazon Mechanical Turk with payment of $0.05 per annotated task per 5 reviewers. The payment of $0.05 was intended to be small enough that spammers and bad users would not feel incentivized to accept the task but was also the average payment per task at the time of submission. With a total of 125 annotations, the total payment to AMT was around $7.00, including miscellaneous fees. The annotators were limited by the percent of previous tasks that have been accepted/approved at a rate of 75% or greater. Assuming that all annotators are fluent English speakers, no other limitations were deemed necessary for this task. There was also no need for master raters or annotators with some form of advanced knowledge.

Figure 5: Setting up the AMT task

[1] Hawkins, Joy; Yelp vs Google: How they deal with fake reviews



Describe your task to Workers

Title Sentiment analysis on yelp reviews

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 label yelp reviews on positivity

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

Keywords sentiment, text, restaurant, yelp

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 5

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

Time allotted per assignment 3 Hours

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.

Worker requirements

Require that Workers be Masters to do your tasks (Who are Mechanical Turk Masters?)

○ Yes ● No

Specify any additional qualifications Workers must meet to work on your tasks:

HIT Approval Rate (%) for all Requesters' HITs | greater than | 75 Remove

(up to 4 more)

(Premium Qualifications incur additional fees, see Pricing Details to learn more)

Project contains adult content (See details)

□ This project may contain potentially explicit or offensive content, for example, nudity.

Task Visibility (What is task visibility?)

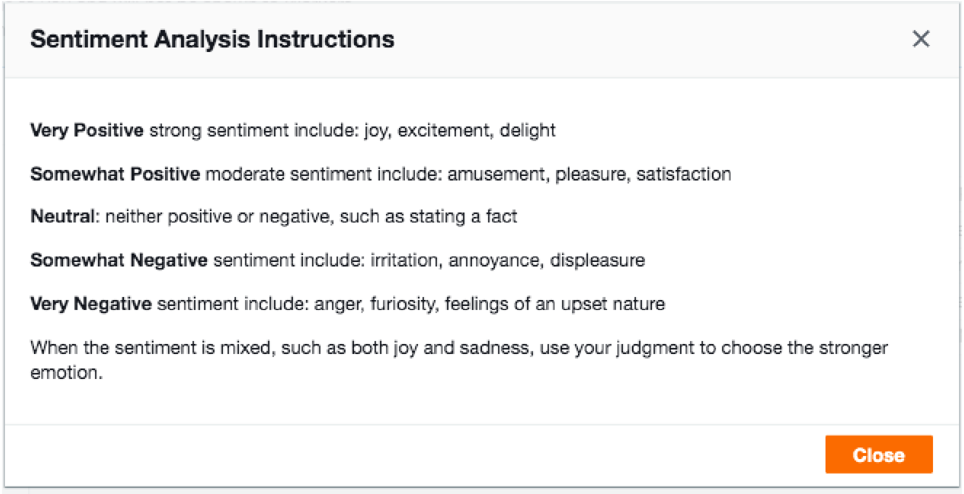
● Public - All Workers can see and preview my tasks

Figure 6: Adding a constraint of HIT acceptance > 75% to ensure good annotations

All five annotators were shown every review in the 25 yelp review set and were asked to label the review with a number from 1 to 5 indicating whether the review was Very Positive, Somewhat Positive, Neutral, Somewhat Negative, or Negative.

These labels were chosen based on personal work experience with in annotation. At a previous annotation project at Yahoo, we found that if the labels were too extreme, it would cause unintended bias. On the project's initial launch, we asked annotators to label ads on relevancy with "perfect", "excellent", "good", "bad", or "poor". We found that regardless of age, gender, professional or educational background, annotators refrained from using the "perfect" label. The reasoning was that from an entire dataset, there should only be one or two cases of perfect relevancy. Our results would always show a maximum of two "perfect" labels and a majority of "good" labels. We re-ran that same exact dataset with different labels that have less of a conceptual bias: these words were "relevant", "somewhat relevant", "barely relevant", "not relevant", and "not applicable" to cover all data issues or other unforeseen issues. We found that annotators began to give a much more normalized distribution of labels over the exact same dataset.

[1] Hawkins, Joy; Yelp vs Google: How they deal with fake reviews



Sentiment Analysis Instructions

Very Positive strong sentiment include: joy, excitement, delight

Somewhat Positive moderate sentiment include: amusement, pleasure, satisfaction

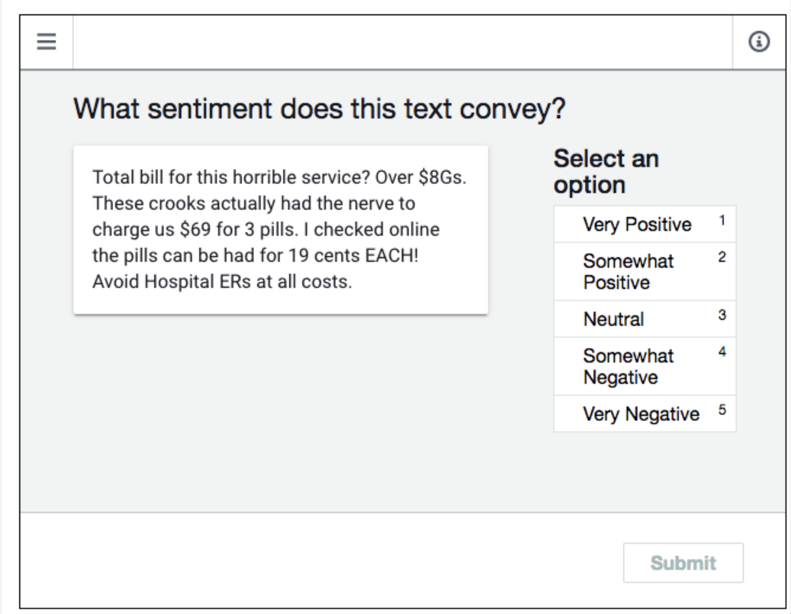
Neutral: neither positive or negative, such as stating a fact

Somewhat Negative sentiment include: irritation, annoyance, displeasure

Very Negative sentiment include: anger, furiosity, feelings of an upset nature

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

Figure 7: This image displays the description of each label chosen for the annotations



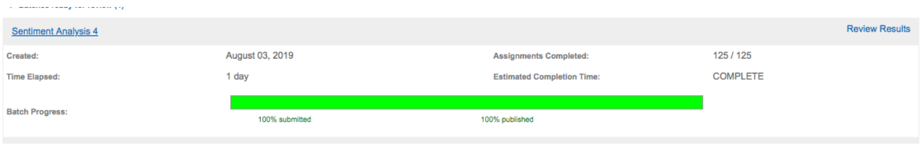
What sentiment does this text convey?

Total bill for this horrible service? Over $8Gs. These crooks actually had the nerve to charge us $69 for 3 pills. I checked online the pills can be had for 19 cents EACH! Avoid Hospital ERs at all costs.

Figure 8: This image displays a sample task/HIT an annotator would see. Each HIT includes a question to answer, a text snippet, and a label to choose for each task.

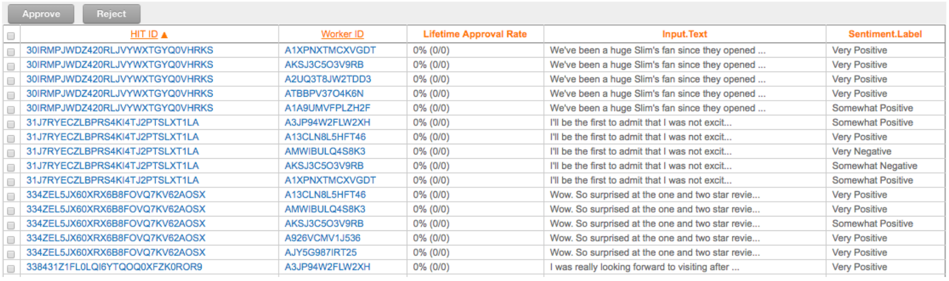
The dataset was submitted to Amazon Mechanical Turk with a time limit of 3 hours for each annotator; annotators were able to complete the dataset in a single day. Each of 5 annotators annotated 25 tasks, which resulted in 125 ratings overall. Amazon Mechanical Turk compiled all of these ratings by task ID and provided a downloadable csv file for further analysis.

[1] Hawkins, Joy; Yelp vs Google: How they deal with fake reviews



[Sentiment Analysis 4]			Review Results
Created:	August 03, 2019	Assignments Completed:	125 / 125
Time Elapsed:	1 day	Estimated Completion Time:	COMPLETE
Batch Progress:	100% submitted 100% published		

Figure 9: status bar for AMT task



□	HIT ID ▲	Worker ID	Lifetime Approval Rate	Input.Text	Sentiment.Label
□	30IRMPJWDZ420RLJVYWXTGYQ0VHRKS	A1XPNXTMCXVGDT	0% (0/0)	We've been a huge Slim's fan since they opened ...	Very Positive
□	30IRMPJWDZ420RLJVYWXTGYQ0VHRKS	AKSJ3C5O3V9RB	0% (0/0)	We've been a huge Slim's fan since they opened ...	Very Positive
□	30IRMPJWDZ420RLJVYWXTGYQ0VHRKS	A2UQ3T8JW2TDD3	0% (0/0)	We've been a huge Slim's fan since they opened ...	Very Positive
□	30IRMPJWDZ420RLJVYWXTGYQ0VHRKS	ATBBPV37O4K6N	0% (0/0)	We've been a huge Slim's fan since they opened ...	Very Positive
□	30IRMPJWDZ420RLJVYWXTGYQ0VHRKS	A1A9UMVFPLZH2F	0% (0/0)	We've been a huge Slim's fan since they opened ...	Somewhat Positive
□	31J7RYECZLBPRS4KI4TJ2PTSLXT1LA	A3JP94W2FLW2XH	0% (0/0)	I'll be the first to admit that I was not excit...	Somewhat Positive
□	31J7RYECZLBPRS4KI4TJ2PTSLXT1LA	A13CLN8L5HFT46	0% (0/0)	I'll be the first to admit that I was not excit...	Very Positive
□	31J7RYECZLBPRS4KI4TJ2PTSLXT1LA	AMWIBULQ4S8K3	0% (0/0)	I'll be the first to admit that I was not excit...	Very Negative
□	31J7RYECZLBPRS4KI4TJ2PTSLXT1LA	AKSJ3C5O3V9RB	0% (0/0)	I'll be the first to admit that I was not excit...	Somewhat Negative
□	31J7RYECZLBPRS4KI4TJ2PTSLXT1LA	A1XPNXTMCXVGDT	0% (0/0)	I'll be the first to admit that I was not excit...	Somewhat Positive
□	334ZEL5JX60XRX6B8FOVQ7KV62AOSX	A13CLN8L5HFT46	0% (0/0)	Wow. So surprised at the one and two star revie...	Very Positive
□	334ZEL5JX60XRX6B8FOVQ7KV62AOSX	AMWIBULQ4S8K3	0% (0/0)	Wow. So surprised at the one and two star revie...	Very Positive
□	334ZEL5JX60XRX6B8FOVQ7KV62AOSX	AKSJ3C5O3V9RB	0% (0/0)	Wow. So surprised at the one and two star revie...	Somewhat Positive
□	334ZEL5JX60XRX6B8FOVQ7KV62AOSX	A926VCMV1J536	0% (0/0)	Wow. So surprised at the one and two star revie...	Very Positive
□	334ZEL5JX60XRX6B8FOVQ7KV62AOSX	AJY5G987IRT25	0% (0/0)	Wow. So surprised at the one and two star revie...	Very Positive
□	338431Z1FL0LQI6YTQOQ0XFZK0ROR9	A3JP94W2FLW2XH	0% (0/0)	I was really looking forward to visiting after ...	Very Positive

Figure 10: Snippet of results

Cohen's Kappa

Cohen's Kappa is a measurement of inter-rater agreement; this measures how likely it is for two raters to come to the same decision based on true agreement versus whether two raters came to the decision by chance. Kappa values closer to one indicate strong intentional agreement and kappa values closer to 0 indicate agreement due to chance. The Kappa coefficient is thought to be a much better test than raw agreement because factoring in 'chance' can be thought of as a form of normalization. Cohen's Kappa requires data be in a categorical form, so the sentiment labels were not converted to their numeric counterparts.

Results

The pairwise Kappa scores were calculated for each pair of annotators. Annotators 2 and 3 seem to have moderately strong agreement with their labels. Annotators 1 and 2 also have moderately strong agreement. The pairs of Annotators 2 and 4 and annotators 3 and 4 have weak agreement but arguably have more intentional agreement than agreement by chance. Annotator 5 has the least agreement with any other annotator which can imply that this annotator may either have not understood the task, or, was randomly choosing labels to finish the task and get paid quicker.

The overall average agreement for all the annotators is 0.236. Surprisingly, despite annotator 5 having the least agreement with the rest of the annotators, removing annotator 5's labels will not significantly increase the average kappa score. The average kappa score without including annotator 5 is only 0.239.

	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5
Annotator 1		0.382	0.170	0.064	0.174

[1] Hawkins, Joy; Yelp vs Google: How they deal with fake reviews

Annotator 2	0.382		0.460	0.314	0.155
Annotator 3	0.170	0.460		0.284	0.145
Annotator 4	0.064	0.314	0.284		0.215
Annotator 5	0.174	0.155	0.145	0.215	

Figure 11: confusion matrix of Cohen's Kappa inter-rater agreement on a scale between 0 and 1

Average kappa score = 0.236

Average kappa (with annotator 5 removed) = .239

Conclusion

While it may be quick and very convenient to crowd source annotations, it is not without risk. Annotators need to be audited to make sure that labels are correct and valid. One method of checking agreement between annotators is to use Cohen's Kappa which will calculate the percent of true agreement while also taking into consideration the probability of agreement due to chance.

Human annotators fare better than machine learning algorithms in that they are able to process nuances more easily; for example: a model may understand the phrase "this is awfully good" as a negative statement whereas a human annotator can more easily overlook the negation to understand the positive intent. Human annotation can be costly, however, so the most optimal method for model training may be a split between algorithmic prediction and human annotation for validation.

Poor labeling will ultimately affect both the restaurant owners and potential customers. Assuming that companies like Yelp use this form of annotation to rank/remove false reviews, if the annotation data is not correct, the algorithm will not properly flag all the negative reviews and may also incorrectly flag positive reviews for removal. This will result in a poor portrayal of the restaurant and can discourage future clientele from patronizing the restaurant.

[1] Hawkins, Joy; Yelp vs Google: How they deal with fake reviews