1
Experiment

1.1 Experimental Design

Note: All experimentation was conducted with IRB approval.

We conducted a survey to determine a baseline of human scores for passages by affect. There were two versions of the survey; both had the same first twelve passages, and a different set of the latter twelve passages. The first twelve passages were gathered by a sentence generator and served as unaltered text. The second set of twelve passages was gathered by feeding the first twelve through one of our replacement models; one version of the survey had all passages go through the arbitrary replacement model, and the other had all passages go through the targeted replacement model. For each passage, we asked the participants to give a score from one to ten in each of the following categories: how natural the passage sounded, how angry the passage sounded, how sad the passage sounded, how joyful the passage sounded, and how fearful the passage sounded.

This was conducted over the internet via Google Forms. As Forms does not support A/B testing, we embedded the surveys in a webpage using a microframework I built to programmatically provide one of the two versions with equal probability. We received 30 responses across the two surveys, with an even split of 15 responses for each version.

1.2 Results

We conducted an independent samples t test on the control results for each category (scores for natural, anger, sadness, joy, fear) to gain insight on the difference between the two groups. Each category has a null hypothesis of the two groups having equal means; each category has the same 12 prompts, and we run the t test on each prompt. The "natural" category had one out of twelve p values below our $\alpha = 0.05$, thus in all other cases we fail to reject the null hypothesis—we are unable to conclude that the two groups are significantly different. Similarly, the "anger" category had one out of twelve, "sadness" had two out of twelve, "joy" had one out of twelve, and "fear" had two out of twelve p values less than $\alpha = 0.05$.

1.2.1 Comparison of groups

Neutral Questions

question	
01.	When motorists sped in and out of traffic, all she could think of was those in need of a transplant.
02.	He drank life before spitting it out.
03.	The toy brought back fond memories of being lost in the rain forest.
04.	Italy is my favorite country; in fact, I plan to spend two weeks there next year.
05.	The blinking lights of the antenna tower came into focus just as I heard a loud snap.
06.	I love bacon, beer, birds, and baboons.
07.	She saw the brake lights, but not in time.
08.	They say that dogs are man's best friend, but this cat was setting out to sabotage that theory.
09.	The tart lemonade quenched her thirst, but not her longing.
10.	He was surprised that his immense laziness was inspirational to others.
11.	They got there early, and they got really good seats.
12.	You can't compare apples and oranges, but what about bananas and plantains?

question	p value	mean difference	upper CI	lower CI	df	Cohen's d	effect size
1.	0.48079	0.57143	1.28675	-0.14389	26	0.27037	small
2.	0.94782	0.07143	0.13751	0.00535	26	0.02498	small
3.	0.34479	0.71429	1.67654	-0.24796	26	0.3637	medium
4.	0.72381	0.28571	0.64294	-0.07151	26	0.13502	small
5.	0.49097	-0.5	-1.19866	0.19866	26	-0.26407	small
6.	0.71533	0.28571	0.65443	-0.083	26	0.13936	small
7.	0.50865	-0.42857	-1.09875	0.2416	26	-0.2533	small
8.	0.37381	-0.57143	-1.47636	0.3335	26	-0.34203	small
9.	0.42139	-0.78571	-1.60264	0.03122	26	-0.30877	small
10.	0.70986	0.35714	0.73329	-0.01901	26	0.14217	small
11.	**0.03027**	1.57143	3.86317	-0.72031	26	0.8662	large
12.	0.30227	0.85714	1.90961	-0.19533	26	0.3978	medium

question	p value	mean difference	upper CI	lower CI	df	Cohen's d	effect size
1.	0.14972	1.21429	2.69871	-0.27014	26	0.56106	medium
2.	0.07658	-1.57143	-3.41563	0.27277	26	-0.69704	small
3.	0.65964	-0.14286	-0.58836	0.30265	26	-0.16838	small
4.	0.45892	-0.14286	-0.89467	0.60895	26	-0.28416	small
5.	0.13268	-0.64286	-2.19515	0.90944	26	-0.58671	small
6.	0.06944	-0.28571	-2.17944	1.60801	26	-0.71576	small
7.	0.8948	-0.07143	-0.20497	0.06211	26	-0.05047	small
8.	0.45384	-0.64286	-1.40329	0.11757	26	-0.28742	small
9.	0.53799	-0.28571	-0.90982	0.3384	26	-0.23589	small
10.	0.18126	-0.42857	-1.80229	0.94514	26	-0.51922	small
11.	0.14574	-0.71429	-2.214	0.78543	26	-0.56684	small
12.	**0.04879**	-0.85714	-2.9245	1.21022	26	-0.78139	small

question	p value	mean difference	upper CI	lower CI	df	Cohen's d	effect size
1.	0.45453	-0.5	-1.25926	0.25926	26	-0.28697	small
2.	0.66341	-0.35714	-0.79737	0.08308	26	-0.16639	small
3.	0.05728	-1.78571	-3.77509	0.20366	26	-0.75191	small
4.	0.14936	-0.21429	-1.70007	1.2715	26	-0.56157	small
5.	0.68706	0.21429	0.62166	-0.19309	26	0.15397	small
6.	0.7837	-0.07143	-0.34878	0.20592	26	-0.10483	small
7.	0.92345	0.07143	0.16845	-0.02559	26	0.03667	small
8.	0.06638	-0.71429	-2.63061	1.20204	26	-0.7243	small
9.	0.08258	-1.28571	-3.09128	0.51985	26	-0.68244	small
10.	1.0	0.0	0.0	0.0	26	0.0	small
11.	**0.03373**	-0.78571	-3.02749	1.45606	26	-0.84731	small
12.	**0.03594**	-0.57143	-2.78369	1.64084	26	-0.83616	small

Natural Scores

Anger Scores

Sadness Scores

Joy Scores

Fear Scores

1.2.2 Error Analysis

Then, we chose a subset of scoring metrics (raw score, and LDA net scores for input and topic)

question	p value	mean difference	upper CI	lower CI	df	Cohen's d	effect size
1.	0.15908	-0.64286	-2.09261	0.8069	26	-0.54796	small
2.	0.91834	0.07143	0.17495	-0.0321	26	0.03913	small
3.	0.71771	-0.35714	-0.72263	0.00834	26	-0.13814	small
4.	0.73313	-0.21429	-0.55893	0.13036	26	-0.13026	small
5.	0.81512	0.07143	0.30764	-0.16478	26	0.08928	small
6.	0.86097	-0.14286	-0.31974	0.03403	26	-0.06686	small
7.	**0.03103**	-0.28571	-2.56607	1.99464	26	-0.86189	small
8.	0.57205	0.5	1.07228	-0.07228	26	0.2163	small
9.	0.06454	1.07143	3.00174	-0.85889	26	0.72959	large
10.	0.42582	-0.71429	-1.52335	0.09478	26	-0.3058	small
11.	0.93493	-0.07143	-0.15386	0.01101	26	-0.03116	small
12.	0.52145	0.64286	1.29277	-0.00705	26	0.24564	small

question	p value	mean difference	upper CI	lower CI	df	Cohen's d	effect size
1.	0.64988	0.35714	0.81638	-0.1021	26	0.17358	small
2.	0.7244	0.21429	0.57071	-0.14214	26	0.13471	small
3.	0.75466	0.21429	0.53011	-0.10154	26	0.11937	small
4.	**0.03103**	-0.28571	-2.56607	1.99464	26	-0.86189	small
5.	0.325	-0.85714	-1.86036	0.14608	26	-0.37918	small
6.	0.13199	-0.28571	-1.8409	1.26947	26	-0.58781	small
7.	0.10962	-1.5	-3.15662	0.15662	26	-0.62614	small
8.	0.86424	-0.14286	-0.31553	0.02982	26	-0.06526	small
9.	0.36909	-0.35714	-1.27119	0.5569	26	-0.34548	small
10.	0.61713	-0.21429	-0.72027	0.2917	26	-0.19124	small
11.	**0.02071**	-0.5	-2.96306	1.96306	26	-0.93095	small
12.	0.21367	-0.28571	-1.56047	0.98904	26	-0.48181	small

means of analysis. We will look at relative error averaged by category, and then subdivided by target and group.

LDA Input Score Error

anger error	sadness error	joy error	fear error	mean error	std dev
-0.85802	-0.76072	-0.70641	-0.79792	-0.78077	0.055211

LDA Topic Score Error

anger error	sadness error	joy error	fear error	mean error	std dev
-0.53525	-0.46815	-0.36493	-0.47174	-0.46002	0.061043

anger error	sadness error	joy error	fear error	mean error	std dev
-0.85647	-0.85213	-0.85312	-0.8497	-0.85286	0.002430

Raw Score Error

We can see that the LDA net scores for topic has the least average error across the board, followed by the LDA input scores, with the raw score trailing at the end. This suggests that the LDA topic score may be the most appropriate scoring model that we offer, although it has the highest variance among error scores. We can further look at our results by group and affect target.

LDA Input Score Error

group	target	anger err	sadness err	joy err	fear err	mean |err|	std dev
neutral		-0.96974	-0.68386	-0.84701	-0.77157	0.81804	0.12113
control		-0.95562	-0.95015	-0.97787	-0.79765	0.92032	0.08266
experimental	anger	-0.43843	-0.48738	-0.20712	-0.98126	0.52855	0.32562
experimental	sadness	-0.56263	0.03343	0.67848	-0.69709	0.49291	0.31204
experimental	joy	-1.0	-1.0	-0.39131	-0.47553	0.71671	0.32892
experimental	fear	-0.79571	-0.94346	-0.89708	-0.9416	0.89446	0.06924

Here, we see that there the LDA input scoring has rather high error with the unaltered (neutral) and random-replacement (control) scores ($> 80\%$), and a mix of high and low error among the the targeted replacement scores. We see high average error within the sentences with a target of "fear" and "joy" (89% and 72% respectively), with lower mean error for sentences targeting "anger" and "sadness" (53% and 49% respectively).

LDA Topic Score Error

group	target	anger err	sadness err	joy err	fear err	mean |err|	std dev
neutral		-0.51351	-0.65078	-0.49961	-0.39885	0.51569	0.10354
control		-0.67275	-0.45844	-0.48453	-0.49215	0.52697	0.09826
experimental	anger	-0.0917	0.13595	0.41926	-0.81972	0.36666	0.33509
experimental	sadness	-0.67467	-0.35851	0.1114	0.45561	0.40005	0.23349
experimental	joy	-0.93078	-0.91615	-0.88804	-0.6649	0.84997	0.12465
experimental	fear	-0.5083	-0.62584	-0.64535	0.08332	0.46571	0.26201

We see that the LDA topic scores are more consistent, largely hovering close to $40 - 50\%$, with sentences that targeted "joy" having much higher mean error at $\approx 85\%$.

group	target	anger err	sadness err	joy err	fear err	mean |err|	std dev
neutral		-0.9763	-0.91245	-0.96795	-0.89303	0.93743	0.04098
control		-0.9187	-0.96874	-0.98527	-0.83505	0.92694	0.06748
experimental	anger	-0.51785	-0.48804	-0.3838	-0.91828	0.57699	0.23467
experimental	sadness	-0.80069	-0.43176	-0.35877	-0.29939	0.47265	0.22529
experimental	joy	-1.0	-1.0	-1.0	-0.72623	0.93156	0.13689
experimental	fear	-0.5926	-0.8712	-0.80598	-0.93233	0.80053	0.14791

Raw Score Error

The raw score model gives scores rather similar to that of the LDA input scores, having high error ($> 80\%$) for the neutral and control groups as well as the sentences targeting "joy" and "fear," with lower levels of error ($40 - 60\%$) in the sentences targeting "anger" and "sadness."

2
Discussion

As per the relative error for each scoring model, we conclude that the LDA topic-scoring model gives the most human-like affect scores. Across all models, we saw a higher error for sentences targeting "joy" and "fear" as opposed to sentences targeting "anger" and "sadness." This may be the result of issues with attempts at changing an underlying tone through word-choice alone, suggesting that theme dominates word-choice in human conceptions; it may also arise from limitations of our lexicons. There may be improvements by constructing our LDA model differently; a comparison between our current model and one based off a large corpus of pre-scored sentences may give more insight as to the appropriateness of our model—this has not been done partially due to the cost of gathering large-scale human data.

2.1 Future Work

Given both the state of the world at the end of this academic year and the scope of this project, there is material left undone. This section serves to address that, both in the sense of acknowledgment and with ideas for implementation if applicable.

There ought to be some LDA replacement model, perhaps targeting a topic with a score closest to the desired output; this could be done by preferring synonyms with high impact on the target topic. Some LDA playground feature could be interesting, such that a user could pass a corpus

and vocabulary to receive a model; this would entail a greater degree of automation as well as some database for models, perhaps with some authentication system for saving remotely. Perhaps our model API should maintain a database connection in order to make less HTTP requests. A fuller featured frontend with more information in regards to analysis is in order; a user should be able to see the effect of each word on the output, and to be able to edit input manually with dropdown menus for word suggestions (based on synonyms). Despite having made a REST API as our CRUD wrapper, I think something like GQL may be more fitting so that we can have finer grained control over the requests. The output from the replacement models should have an option to score the generated text directly as opposed to copy-pasting the output, and on a similar note, we should have an option to save results for reference (via `localStorage`). Then, there should be several other UI niceties: light/dark mode should extend to fully support things like the documentation and the info popovers, there should be more info popovers across the site, we should suggest words in the thesaurus page if no results exist (via edit distance), and—for development—there should be some UI playground for mocking designs (this can be done by registering components globally, providing a textarea for the template, and compiling/rendering the results).