Sentiment Analysis: Using ML to quantify public sentiment on people, topics, or organizations

Cody Houff
Georgia Tech
ME 8813 Machine Learning

1: Introduction

In the last decade there has been a large growing collection of opinions on the Internet, and especially on social media. This can be quite useful for policymakers, individuals, or organizations to understand the needs and problems of societies and formulate effective strategies for addressing them. Today many individuals argue there exists a tangible disconnect of various companies, figure heads, and politicians with the general population. In particular, the younger generation seems to be particularly disconnected. This is evident when browsing social networking sites such as Facebook, Instagram, Twitter, Reddit and reading the discussions. Quantifying this sentiment is difficult using traditional means. How does one quantify approval or disapproval from social media or news sources? Traditional polling methods or study groups can be slow, expensive, and sometimes ostracize a portion of the population. For example, calling only voters or customers with landlines can result in a poor sample of likely because some demographic groups have few landlines. Monitoring social networks represents a potential positive addition to the above methods. This method for capturing people's opinions overcomes the low-response rate problem and other problems that can arise from polling. People that use social networks naturally express their preferences in online discussions without being exposed to direct questions. This data can be collected with machine learning through sentiment analysis. Sentiment analysis is the use of natural language processing to systematically identify, extract, quantify, and study subjective information. My hope is that through deep learning sentiment analysis policymakers, individuals, or organizations can more effectively take calculated actions and create policies in line with public sentiment.

2: Methodology

Part 1

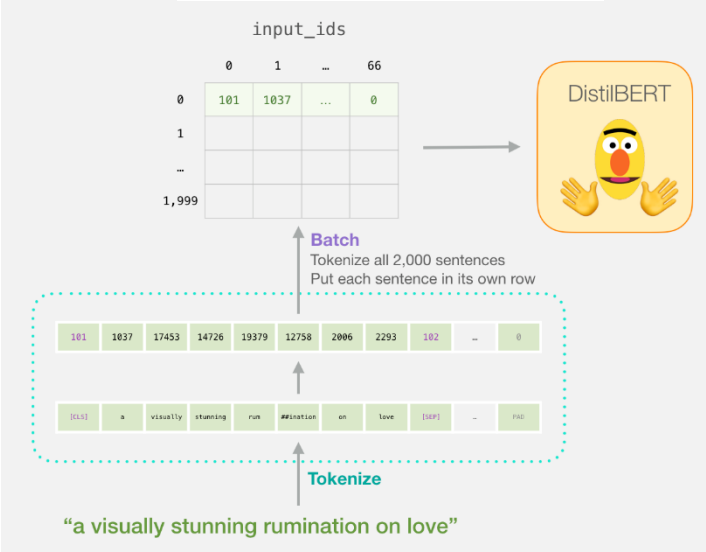
There were two main goals to this research: the part 1 was designing, coding, training a sentiment analysis model that is around 80% to 95% accurate. The second goal was to use a very large existing pretrained language model on a reddit and twitter dataset to find insides on a topic. For the first goal, there a few main steps I took to doing sentiment analysis using deep learning. The first step was to find a high quality labeled dataset to train and test on. For training and testing I used the popular labeled dataset Stanford Sentiment Treebank SST-2 [1]. This dataset contains 215,154 phrases with fine-grained sentiment labels in the parse trees of 11,855 sentences from movie reviews. Each sentence is labeled with a 1 or 0 corresponding to a positive or negative sentiment shown below (Figure 1).

Figure 1: SST-2 dataset training

	Reviews	Ratings
0	a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
1	apparently reassembled from the cutting room floor of any given daytime soap	0
2	they presume their audience wo n't sit still for a sociology lesson , however entertainingly presented , so they trot out the conventional science fiction elements of bug eyed monsters and futuristic women in skimpy clothes	0
3	this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
4	jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1
...	...	...
4995	just about the best straight up , old school horror film of the last 15 years	1
4996	in the director 's cut , the film is not only a love song to the movies but it also is more fully an example of the kind of lush , all enveloping movie experience it rhapsodizes	1
4997	samuel l jackson is one of the best actors there is	1
4998	it does give a taste of the burning man ethos , an appealing blend of counter cultural idealism and hedonistic creativity	1
4999	a plethora of engaging diatribes on the meaning of ` home , ' delivered in grand passion by the members of the various households	1

The next step is formatting the data to be used is the transformer. I used a Tokenizer which is the process of converting text into tokens before transforming it into vectors of numbers shown below (Figure 2). After that I needed to add some padding so that all the inputs are the same size.

Figure 2: Tokenizer diagram



I then feed in the training and test data into a transformer. More specifically a type of transformer called BERT, which stands for Bidirectional Encoder Representations from Transformers (BERT) [2]. It is a way of learning representations of a language that uses a transformer, specifically, the encoder part of the transformer. In the BERT paper the base model is a 12-layer, 768-hidden, 12-heads, 110M parameter neural network architecture. I was able to get good results using an architecture similar to DistilBERT (6-layer, 768-hidden, 12-heads, 66M parameters) [3]. Below is a basic diagram of BERT (Figure 3). The large models discussed can also be compared in the table below (Figure 5). Lastly I used logistic regression. Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. Logistic regression equation shown below (Figure 4).

Figure 5: Models Compared

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

Figure 3: BERT

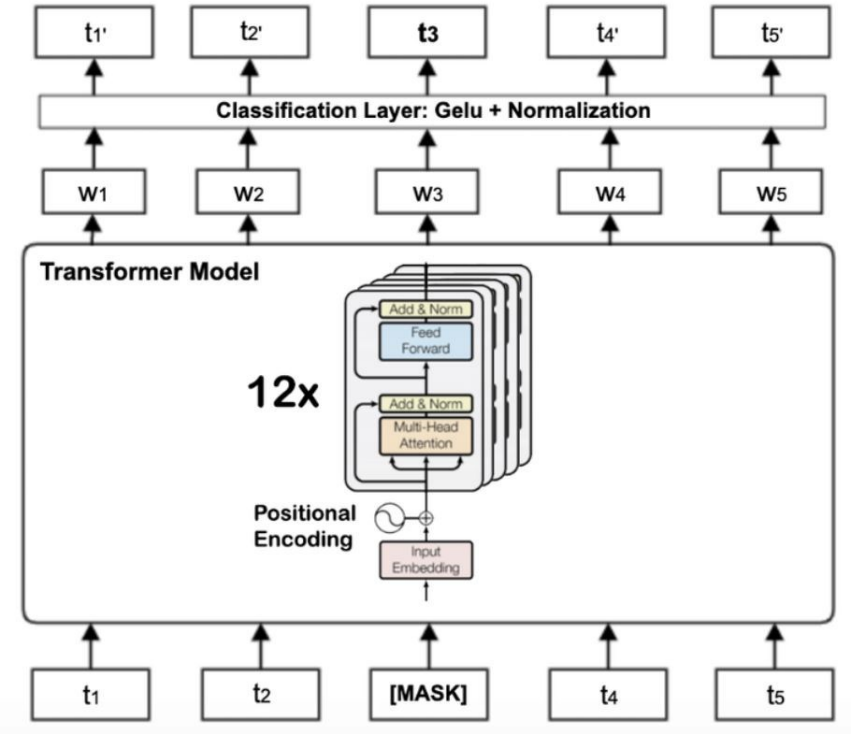


Figure 4: Logistic Regression

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Part 2

For part 2 I wanted to use a very large pretrained model on data from reddit and twitter. The topic I was going to collect data on and do analysis on with machine learning was Elon Musk taking over twitter. My hope is that with enough data from twitter and reddit I can quantify how people feel about the change in leadership. Thankfully there is an API in place for both twitter and reddit so there is no need for time consuming web scraping. Researchers need to create a developer account to get access with their 2 individualized keys. For twitter it was less straight forward. There is a lengthy process of creating developer account and submitting a request explaining why you want access to the API. There are different tiers of access you get with twitter: Essential, Elevated, and Academic Research. I was able to get access to Elevated which allowed me to collect data in a semi-limited way. I was able to collect data by saving tweets using the key words “Elon + Musk +Twitter”. I accessed a limited number of tweets from each

hour of each day for the past few days. Overall I collected almost 14,000 tweets associated with the above keywords and stored them in a csv file. I then cleaned the dataset so it could be tokenized for the model. For reddit, I looked at the top 250 upvoted posts in the last week with the keywords being “Elon + Musk +Twitter” [18]. Then I collected to top 20 upvoted comments in each of those posts and stored them in a csv file. It ended up being totaling about 5200 comments. Similarly to the twitter dataset I cleaned it so it could be properly tokenized.

The large pretrained Model I chose to use is called RoBERTa (figure 5) which is a robustly optimized method for pretraining natural language processing (NLP) systems that improves on Bidirectional Encoder Representations from Transformers [16]. I looked at a specific RoBERTa model trained on a twitter dataset [17]. I was able to download the model and similarly to my costume model it predicts with a confidence percentage if the sentence is positive or negative. Only difference with RoBERTa is it also predicts neutral sentiment in addition to negative or positive. I tested the pretrained model on the two datasets I created.

3: Results

Part 1

My results with the custom model on the Stanford Sentiment Treebank SST-2 using 5000 sentences with a train test split of 80/20 was around 86% accuracy (Figure 5). This means 86% of the time the model can accurately predict if a opinion or statement was positive or negative. This is surprisingly good as I was limited with my personal computer in terms of training. I am confident I can improve that accuracy with more training data and tuning the model parameters however this will definitely require a much more powerful computer system as mine was pushed

to its limit. For comparison state of the art models (that are much larger) trained with multiple top of the line GPUs has an accuracy of around 95%. The sentences it had most trouble with seemed to be the ones with mixed sentiment ending in negative. Also, sentences that were sarcastic or joking in nature.

Figure 5: Custom Model Predictions (86%)

	Reviews	Ratings	Prediction
0	no movement , no yuks , not much of anything	0	0
1	a gob of drivel so sickly sweet , even the eager consumers of moore 's pasteurized ditties will retch it up like rancid cr me br l e	0	0
2	gangs of new york is an unapologetic mess , whose only saving grace is that it ends by blowing just about everything up	0	1
3	we never really feel involved with the story , as all of its ideas remain just that abstract ideas	0	0
4	this is one of polanski 's best films	1	1
...	...	...	...
1816	an often deadly boring , strange reading of a classic whose witty dialogue is treated with a baffling casual approach	0	0
1817	the problem with concept films is that if the concept is a poor one , there 's no saving the movie	0	0
1818	safe conduct , however ambitious and well intentioned , fails to hit the entertainment bull 's eye	0	0
1819	a film made with as little wit , interest , and professionalism as artistically possible for a slummy hollywood caper flick	0	0
1820	but here 's the real damn it is n't funny , either	0	0

Part 2

For the part 2 results were very interesting to say the least. And again the test data was collected by searching key words “Elon + Musk +Twitter” to get a sense of how people feal about the change in leadership. Specifically, for reddit it was the top 20 comments of the 250 posts related to the key words over the last week. For reddit the results were out of 5187 comments, 2310 were negative, 2349 were neutral, and 528 were positive (Figure 6). There were almost 4.5x more negative comments than positive. So it seems obvious from the reddit community that most are not happy with Elon taking over. To better visualize the kinds of things people are saying I created a word cloud (Figure 7). Also shown below are the results table with confidence

percentages (Figure 8). What we can deduce from this is twitter may feel some push back or people leaving the platform with the reddit crowd if sentiment stays like this.

Figure 6: Top Reddit Comments Sentiment

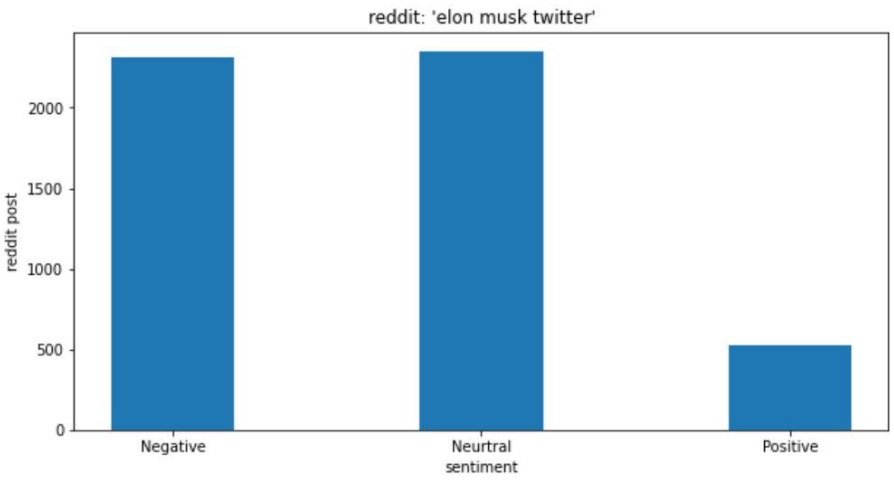


Figure 7: Top Reddit Comments WordCloud

Figure 8: Top Reddit Comments Results Table

	Text	Negative	Neutral	Positive
0	FORTY FIVE BILLION jesus well I guess Im rich in spirit at least	0.0136	0.1074	0.8790
1	Between Netflix making crazy choices and now this Im almost wondering if we are seeing the Dotcom bubble 20 about to burst somehow	0.6965	0.2860	0.0175
2	I hope he buys it just for it to become the MySpace	0.2841	0.5861	0.1299
3	nan	0.0936	0.6809	0.2255
4	nan	0.0936	0.6809	0.2255
...	...	...	...	...
5182	The one time I caught a suspension I was talking about Blackrock	0.2868	0.6753	0.0379
5183	I want to see blackfacehitler trend again	0.3387	0.5584	0.1029
5184	If thats whats trending thinking it is trending so vangaurd doesnt	0.2570	0.7067	0.0363
5185	Financial Markets giant trending during one of the most volatile weeks in history You dont say	0.6043	0.3648	0.0310
5186	The deal hasnt gone through yet though right Hes not yet the owner	0.1091	0.8704	0.0205

5187 rows × 4 columns

The results for twitter posts were noticeably less polarizing with 3856 negative, 7876 neutral, and 2000 positive (Figure 9). Although there were much less polarizing negative posts about Elon taking over, there were still almost 2x as many negative posts as positive. I also included the table of results (Figure 11) and a word cloud for visualizing the most commonly used words

they used (Figure 10). Overall, the sentiment between reddit and twitter seems to be pretty negative. Reddit had much more negative things to say than twitter and twitter is noticeably more neutral on the subject compared to reddit.

Figure 9: Twitter Posts Sentiment

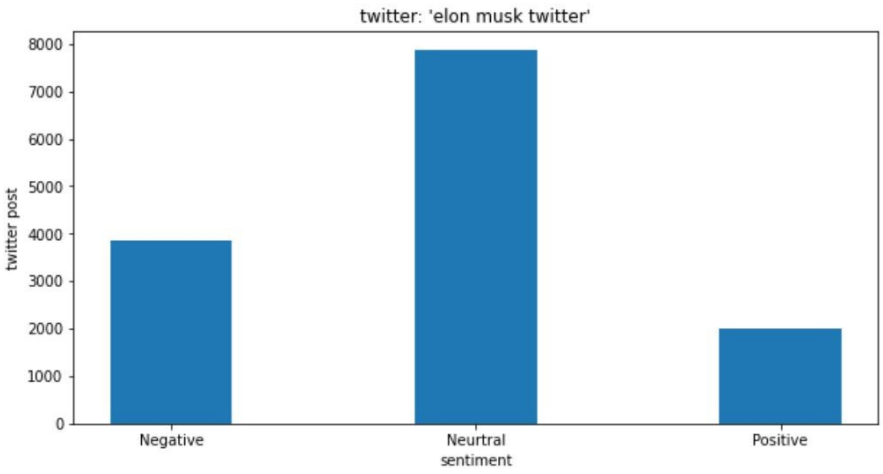


Figure 10: Twitter Posts Sentiment

Figure 11: Twitter Posts Results Table

	Text	Negative	Neutral	Positive
0	onlinebusiness digitalmarketing makemoneyfast Elon Musk Posts A Series Of Engaging Tweets Following His 44 Billion Buyout Of Twitter http The world has been at a relative standstill since it was confirmed that Elon Musk was the http	0.0173	0.8245	0.1582
1	BoydCothran one of the Journals editors was recently interviewed for a NBCNews piece on Elon Musks takeover of Twitter http	0.0067	0.9361	0.0573
2	Charlie Kirk returns to Twitter Conservatives celebrate Musk takeover http found this article on MxM Newsa new app without media bias or censorship Download it here http	0.0045	0.5254	0.4701
3	The Washington Posts Repulsive Defense Of Twitter Execs Makes Even Elon Musk Look Goodhttpstcou2spmn6vBS	0.2892	0.6114	0.0994
4	MUSK blocked PUBLIC CITIZEN That does not mean he has banned it from Twitter just from his own account It is still a bad sign if Elon cannot read reasonable progressive thinking from a reasonable source Its an alarming tell	0.8194	0.1732	0.0075
...	...	...	...	...
13727	Twitter is reportedly reconsidering Elon Musks bid to buy the social media company after the Tesla CEO confirmed 465 billion in financing http	0.0212	0.9318	0.0470
13728	BREAKING WSJ Reports Twitter Will Reconsider Elon Musk Offer http	0.0076	0.9174	0.0750
13729	Twitter is reportedly reconsidering Elon Musks bid to buy the social media company after the Tesla CEO confirmed 465 billion in financing http via Yahoo	0.0173	0.9301	0.0526
13730	Say goodbyeElon Musks cryptic moving on tweet sparks speculation hes done with Twitter takeover http SmartNews	0.1599	0.8082	0.0319
13731	Twitter Meeting with Elon Musk to Discuss 43 Billion BuyOut Proposal http via TMZ	0.0106	0.9351	0.0543

4: Summary and Future Work

A possible next step is getting access to twitters higher level API that will allow more open use and more access to tweets and further back in history [14]. I could also try other large models to see if I get similar results. I would also like to look at other ways of visualizing the data. That could look like weighting the result by the confidence percentage. It could look like, coming up with a smart way of combining sentiment analysis with word cloud. Maybe sorting the comments into negative and positive then making the word cloud from each to get an idea why people are positive or negative. Something that would be interesting is turning it into a live tracker that updates every day on public sentiment on people, topics, or organizations that the user specifies. This could be around a social movement such as BLM or political issues such as abortion. It could also track public sentiment on certain government officials, especially someone running for office. This could be something someone check every morning like they do the weather. If someone was pursuing stock trading sentiment analysis like this could be extremely valuable tool to have. A lot of securities on the stock market react to public sentiment especially with certain volatile securities like crypto currencies.

Overall, my research has found that the reaction to Elon Musk taking over twitter has been largely negative. Reddit users had much more negative things to say than twitter users and twitter users were noticeably more neutral on the subject compared to reddit users. Reddit had about 4.5x more negative comments than positive and twitter had almost 2x as many negative posts as positive. However, a potential issue with this research is it fails to take into account bots or fake users. This is mostly an issue with twitter since the reddit posts are sorted by upvotes it is possible to sort out the spam or bots. Twitter has a hug issue with bot accounts so its easy to see

how this could skew the data. Additionally, people with means can and have bought bot accounts to promote or get momentum going around chosen ideas.

Works Cited

[1] Socher, Richard κ.ά. 'Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank'. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* Seattle, Washington, USA: Association for Computational Linguistics, 2013. 1631–1642. Web.

[2] Devlin, Jacob κ.ά. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. 2018. Web.

[3] Sanh, Victor κ.ά. 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter'. 2019. Web.

[4] Pagolu, Venkata Sasank κ.ά. 'Sentiment Analysis of Twitter Data for Predicting Stock Market Movements'. *CoRR* abs/1610.09225 (2016): n. pag. Web.

[5] Kabbani, Taylan, και Fatih Enes Usta. 'Predicting The Stock Trend Using News Sentiment Analysis and Technical Indicators in Spark'. 2022. Web.

[6] Gupta, Shashank. "Sentiment Analysis: Concept, Analysis and Applications." *Medium*, Towards Data Science, 19 Jan. 2018, https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

[7] "Sentiment Analysis Using Python - NewsCatcher." , 18 Jan. 2022, https://newscatcherapi.com/blog/sentiment-analysis-using-python.

[8] Rajapakse, Thilina. "To Distil or Not to Distil: Bert, Roberta, and Xlnet." Medium, Towards Data Science, 7 Feb. 2020, https://towardsdatascience.com/to-distil-or-not-to-distil-bert-roberta-and-xlnet-c777ad92f8.

[9] Ganta, Raviteja. “Sentiment Analysis Using Bert and Hugging Face.” Ravi.ai, 24 Nov. 1970, https://raviteja-ganta.github.io/sentiment-analysis-using-bert-and-hugging-face.

[10] Liu, Yinhan κ.ά. ‘RoBERTa: A Robustly Optimized BERT Pretraining Approach’. *CoRR* abs/1907.11692 (2019): n. pag. Web.

[11] Yang, Zhilin κ.ά. ‘XLNet: Generalized Autoregressive Pretraining for Language Understanding’. CoRR abs/1906.08237 (2019): n. pag. Web.

[12] “Cardiffnlp/Twitter-Roberta-Base-Sentiment-Latest at Main.” Cardiffnlp/Twitter-Roberta-Base-Sentiment-Latest at Main, https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest/tree/main.

[13] Briggs, James. “How to Use the Reddit API in Python.” *Medium*, Towards Data Science, 2 Sept. 2021, https://towardsdatascience.com/how-to-use-the-reddit-api-in-python-5e05ddfd1e5c.

[14] “Twitter Development Portal.” Twitter, Twitter, https://developer.twitter.com/en/portal/dashboard.

[15] Edward, Andrew. “An Extensive Guide to Collecting Tweets from Twitter API V2 for Academic Research Using Python 3.” *Medium*, Towards Data Science, 17 June 2021, https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a.

[16] Liu, Yinhan κ.ά. ‘RoBERTa: A Robustly Optimized BERT Pretraining Approach’. CoRR abs/1907.11692 (2019): n. pag. Web.

[17] “Cardiffnlp/Twitter-Roberta-Base-Sentiment-Latest at Main.” *Cardiffnlp/Twitter-Roberta-Base-Sentiment-Latest at Main*, https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest/tree/main.

[18] “Api Documentation.” Reddit.com: *API Documentation*, https://www.reddit.com/dev/api/.