Garett Ruping

**Project proposal [1 to 2 pages (PDF); Post on Canvas]**

Each group should post a 1-2 page project proposal in PDF to the Canvas. Your proposal should include a brief, descriptive name of your project name. Your name should be something memorable!

In the proposal, you should address the following issues:

* **What is exactly the function of your tool (or a method)? That is, what will it do?**
* **Why would we need such a tool (or a method) and who would you expect to use it and benefit from it?**
* Does this kind of tools/methods already exist? If similar tools/methods exist, how is your tool/method different from them? Would people care about the difference? How hard is it to build such a tool/algorithm? What is the challenge?
* **How do you plan to build it? You should mention the data you will use and the core algorithm that you will implement (either existing algorithm for tools or new algorithm for methods).**
* **What existing resources can you use?**
* **How will you demonstrate the usefulness of your tool/method?**

Possible intro? Good review of why companies are about using sentiment analysis:

<https://www.businessnewsdaily.com/10018-sentiment-analysis-improve-business.html>

## 

## Businesses can profit from finding out [what their target customers want](https://www.businessnewsdaily.com/8714-know-target-customer.html) and think about their company and its products or services in real time by conducting sentiment analysis. Sentiment analytics apps have the potential to revolutionize the [relationship between brands and their consumers](https://www.businessnewsdaily.com/2821-consumers-relationships-brands.html) by creating greater understanding. Businesses can use the data from a sentiment analysis to drive revenue and guide marketing efforts.

## Sentiment analysis provides understanding to companies about how consumers feel and what they want from a business. Businesses need ways of monitoring real-time conversations about their company and its products or services to measure consumer sentiment in order to improve brand reputation. The data from sentiment analysis can be used to determine which products and services their customers want or how they’re feeling about a brand.

Potential models:

1. SVM - Jennifer
2. Bernoulli Naïve Bayes - Emerald
3. Logistic Regression - Emerald
4. BERT - Anson
5. CNN - Lijian
6. RNN - Garett
7. Attention - Garett
8. HW5 - Jennifer <https://arxiv.org/pdf/1810.04805.pdf>
9. Tf-idf analysis

## 

## Businesses can use sentiment analysis to see how well their marketing campaigns are going on social media and third-party websites. With brand-new product launches, they can scan online comments to see if any customers are having issues. Companies can also get a sense of how well their target audience has received their new product. Based on the results of the analysis, they can adjust their sales and marketing plans to feed into or address consumer sentiment. Traditional social media monitoring often focuses on measuring the number of likes, comments and shares a post gets. While these numbers might indicate buzz around a company, they don’t give emotional insights into consumers’ likes, dislikes and expectations. In contrast, companies can use sentiment analysis to understand whether consumers feel ‘positive,’ ‘negative’ or ‘neutral’ about a certain brand, product or topic.

The tool we are creating performs sentiment analysis on tweets. Our plan is to train multiple models on a twitter sentiment dataset off of kaggle ([Sentiment-140 Twitter Dataset](https://www.kaggle.com/datasets/kazanova/sentiment140)). After training them specifically on the twitter dataset, we plan to use the model to also find the sentiment of Elon Musk tweets. We may look into adding data to the dataset if needed, and preprocessing that could help in training. This tool can be useful to quickly identify the sentiment of a large amount of tweets. Users can create a batch of tweets that interest them and use the tool to quickly find the sentiment of individual tweets or the whole batch, not specifically tied to an individual but all of twitter. This can be useful for people and companies that are interested in specific topics or to learn about the sentiment pertaining to products or ideas. There are already different tools that exist such as a naive bayes classifier ([Naive Bayes](https://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf)) as well as SVM approaches ([SVM](https://ieeexplore.ieee.org/abstract/document/7066632?casa_token=_0cERclAgqcAAAAA:6j2QKr3R3GOzmj-tV4f9H7ox37vDKHb2X1TLyNEJxY-GnOSg9MKsR57QpJPqes1qkF18a2_h)). As twitter sentiment analysis is a well studied field, we plan on expanding on these approaches by training new models and comparing how well they perform. The potential models we plan on exploring include: SVM, Bernoulli Naïve Bayes, Logistic Regression, BERT, CNN, RNN, and Attention.

Different from the existing models, the team is aiming to create an ensemble of the best performing models, upon finishing training these models and evaluating their performance. Since an ensemble of different models would be more robust, considering the decision will come from multiple models instead of a single model. Another aspect that is different from the existing model, the team will attempt to create a brand new model which combines the best aspects of each well-performed model, for instance, the team may attempt a neural network that contains both convolutional layers and recurrent layers. It will also be interesting to examine in which order the characteristics should be applied, for instance whether it is better to apply a convolutional layer before the recurrent layer or vice versa.

The main difficulty with this project would be to create and train all the models, as well as creating a unique model based on what we think would be the best parts of the most accurate performing models. Thankfully, the resources that can be helpful to this study are abundant. The resources that will be consulted included but not limited to these articles: [Emotion and sentiment analysis of tweets using BERT](https://www.researchgate.net/publication/350591267_Emotion_and_sentiment_analysis_of_tweets_using_BERT), [Twitter sentiment analysis using deep learning models](https://ieeexplore.ieee.org/document/9342279), [Sentiment Analysis of Twitter Data Using TF-IDF and Machine Learning Techniques](https://ieeexplore.ieee.org/document/9850477), [Enhanced Twitter sentimental analysis using artificial neural network over logistic regression towards increase in accuracy of prediction](https://aip.scitation.org/doi/abs/10.1063/5.0074806?journalCode=apc). Using those papers we can learn how the authors created the models in practice. We can also use google colab and existing machine learning libraries, like tensorflow and pytorch to construct and examine the models. Our end product would be a website that has a live feed of specific user’s tweets and associated sentiment. We would also show statistics of our new models to demonstrate their performance against existing models.

We will train, validate, and test the model with the dataset [Sentiment-140 Twitter Dataset](https://www.kaggle.com/datasets/kazanova/sentiment140) from Kaggle, and upon the completion of this process, the model will be implemented on a different dataset created by the member of the team. The dataset [Sentiment-140 Twitter Dataset](https://www.kaggle.com/datasets/kazanova/sentiment140) includes 1.6 million data points, each data point includes information of a tweet. The target variable is the polarity of the tweet, which includes three categories with 0 indicates negative, 2 indicates neutral, and 4 indicates positive. The features provided by this dataset includes ids which is the unique identifier of a specific tweet; date which provide information regarding the date and time a tweet was posted; flag which shows that whether there is a query, and the user which provides information about the account that posted the tweet; lastly, the text which is the content of the tweet. Out of all the features provided, the only feature that will be useful is the content of the tweet itself because the information such as tweet id, query flags would not be universally available in other datasets. From further observing the target variable, there was no instance of class neutral, therefore the classification would be a straightforward binary classification. To make the process easier to comprehend, the team decided it makes sense to relabel the positive instance as ‘1’ instead of ‘4’. Since the dataset is large enough, the dataset will be split into train, validation, and test set by the ratio of 60%, 20%, and 20%. To ensure the proportion of the positive and negative instance remains the same, the dataset will be split in a stratified manner, this will eliminate the possibility that the training set has most of the one instance, but the validation set and the test set has another. After processing the data, the team will start the experiments and the construction of the model.

Once we finish training the model, we will evaluate the sentiment of 100 most recent Elon Musk’s tweets (without training the model specifically on his tweets). Our team will manually label these tweets and compare that against our model’s prediction to see its accuracy.