PROPOSAL

Graham Harris, Judy Zhang, Christine Ma, Jong Heon Han

PROBLEM STATEMENT

Given a dataset of reviews with an assigned binary classified sentiment of positive and negative, represented in the data set as probabilistic range between 1 and 0 respectively, our team aims to test different tokenization methods on a reputable model in order to analyze how various tokens are weighted in their sentiment value. Several different tokenization methods will be compared to discern the best system.

DATASET

- 50,000 IMDb movie reviews
  - Balanced distribution of positive and negative
  - 25,000 in test set
  - 25,000 in training set
  - 50,000 unlabeled reviews, will not be used during the testing process
  - Web link: https://ai.stanford.edu/~amaas/data/sentiment/.
- Entries in training_set.txt format:
  - Text filenames act as IDs
  - Sentiment labeled as 1 (positive) or 0 (negative)
  - Movie review text
- Entries in test_set.txt format:
  - Same as training set, without labeled sentiment
- Entries in answer_key.txt format:
  - Same as training set, without movie review text
  - *Additional file answer_key_with_review.txt contains the movie reviews*

MODEL AND DIFFERENT TOKENIZATION METHODS

- Model: Linear Regression
  - Input: Tokenized movie reviews

    - Output: Discrete movie sentiment label (0 for negative, 1 for positive)
- Tokenization Methods:
    - Split by space (control)
    - Adjectives as sole tokens
    - Ignoring punctuation
    - TF-IDF weights
    - Stemming included

EVALUATION METHOD

Our team will use recall, precision, and f-measure to evaluate token methods. These results will then be compared to answer_key.txt.

ROLES AND COLLABORATION

1. Jong Heon Han - Data Analyst, Evaluation
2. Graham Harris - Writer, Editor, Researcher
3. Christine Ma - General Coder, Dataset Implementation, Tokenization
4. Judy Zhang - Researcher, Tokenization

Graham and Judy will work together on the research component. Christine and Judy create the tokenizaters. Jong will perform analysis, the results of which will be communicated to the group. Collectively we will analyze the results and determine conclusions.

ACADEMIC ARTICLES

- Breaking Sentiment Analysis of Movie Reviews (link)
    - Movie reviews allow for pragmatic and stylistic manipulations, it is difficult for systems to properly recognize sentiment
- Learning to Generate Reviews and Discovering Sentiment (link)
    - Sensitivity of learned representations of data models on distributions that they are trained on
- Overcoming Language Variation in Sentiment Analysis with Social Attention (link)
    - Group together reviews with similar language patterns

- How to Fine-Tune BERT for Text Classification? (link)
    - Sentiment analysis tasks used to rigorously test different fine-tuning methods of BERT on text classification
    - Yelp reviews proposed for sentiment analysis
- Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings (link)
    - Best results were obtained on untrained data, suggests that there were complementary text characterizations in their model
    - May be beneficial to try to combine multiple tokenization methods for optimal results (if the methods are complementary)

CITATIONS

Bonfil, Ben, and Ieva Staliunaite. "Breaking Sentiment Analysis of Movie Reviews." *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, 2017, pp. 61-64, https://www.aclweb.org/anthology/W17-5410.pdf.

Johnson, Rie, and Tong Zhang. "Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings.", 2016, https://arxiv.org/pdf/1602.02373.pdf.

Raford, Alec, Rafal Josefowicz, and Ilya Sutskever. "Learning to Generate Reviews and Discovering Sentiment.", 2017, https://arxiv.org/pdf/1704.01444.pdf.

Maas, Andrew, et al. *Learning Word Vectors for Sentiment Analysis*. www.aclweb.org/anthology/P11-1015.

Sun, Chi, et al. "How to Fine-Tune BERT for Text Classification?", https://arxiv.org/pdf/1905.05583.pdf.

Yang, Yi, and Jacob Eisenstein. "Overcoming Language Variation in Sentiment Analysis with Social Attention." *Transactions of the Association for Computational Linguistics*, vol. 5, 2017, pp. 295–307, https://www.aclweb.org/anthology/Q17-1021.pdf.

Idea: Keyword Extraction/Classification

1. Choose a dataset
2. As a start everyone (4) chooses a language model (4 total)
    a. Then find a reputable code/source on github
    b. Run it on the data set
    c. Analyze the error percentage
3. Compare to answer key and draw conclusions
4. If time, run more language models
5. Ultimately, present findings

So this is more like a research paper than creating a new project.

Paper Basic Structure can be..

0. Abstract
1.Introduction
2.Task Description
3.Datasets (example: preprocessing, training data,development data...)
4. Language Models
5. Results and Analysis

NOTES FROM TIN:

How do we compare positive and negative?

- Probability distribution:
    - Computer assigns probabilities of negative and positive
    - Output is between 0 and 1 (or 0 and something else)
    - EX: 0 is negative and 1 is positive
- Code classes are models, eval models are classes, etc
    - Keep classes exactly how they are
    - Make sure input and output sizes are known?
- Model is only a line of code
- Scikit-learn (all the models and encoding methods are here), Binary classification, Decision tree, SVM, logistics, sklearn

PICK ONE MODEL, IMPROVE ON IT

FIND VARIETY OF TOKENIZATION METHODS TO FEED TO MODEL

He recommends:

- One person in charge of evaluation, one of model (working together so things run smoothly)
- One person in charge of data set (does it fit in the model?)

MAIN INVESTIGATION: what tokenizers do we want to explore?

- Ex: split by spaces, remove !, remove words that are irrelevant (like just adjectives, etc)
- Does order matter? BERT embedding,

EXPECTATION: Have model by Tuesday, https://www.statlearning.com/, why x embedder is better than y embedder (examples of where the machine is right or wrong)

ROLE BREAKDOWN?

Writer, editor, researcher, coder, analysis, diagrams, GitHub
https://github.com/gwharris/MovieSentimentAnalysis
dataset:
https://ai.stanford.edu/~amaas/data/sentiment/

Model Possibility = Logistic Regression, SVM
Tokenizer = Split by space, adjective tokens, ignore punctuation (!), tf-idf, stemming added
PROPOSAL: (bold means mandatory, non-bold is responses)

- **List of members and proposed roles**
    - Jong Heon Han: Data Analyst, Evaluation
        - precision, f-measure, etc
    - Graham Harris: Writer/Editor
        - Graham & Jong work together on what we are looking for in our end result/analysis
    - Christine Ma: Coding, Tokenizer, DataSet
        - write tokenizers (2), describe what method used and why for the writing team
    - Judy Zhang: Researcher,
        - write tokenizers (2), describe what method used and why for the writing team
- **Problem statement or introduction** (keep short)
    - Given a dataset of movie reviews with an assigned sentiment ranging from negative, neutral, and positive, test different language models and analyse accuracy to draw conclusions regarding which model performs the best in sentiment analysis.
- **Evaluation plan**:
    - **1) Task/output our system will provide**
    - **2) Measure of how well our system did**
        - Accuracies on NLP progress site
        - Precision, recall, f-measure?
        - Compare to state-of-the-art
    - **Discuss procedure to ensure results are valid**
- **Discuss 5 academic articles related to the topic:**
    - **How does each article relate to our proposed research question?**
    - 1) https://arxiv.org/pdf/1810.02840.pdf (https://github.com/HazyResearch/metal)
        - uses weak supervision, which eliminates the problem of getting sufficiently large hand labeled data sets
        - https://www.snorkel.org
        - Repo on github is from 2019, see if we can find 2018 version, or just compare from different years if necessary

    - 2) https://arxiv.org/abs/1704.01444 (https://github.com/NVIDIA/sentiment-discovery)
        - uses a small labeled dataset to achieve state of the art results on the Stanford Sentiment Treebank in 2018
        - Going to compare it to another highly successful method from 2018
    - 3)just put in the most recent most accurate papers from the sentiment analysis section of nlp progress and use them to figure out why the current state of the art is more successful than the two methods we focused on from 2018
    - 4)
    - 5)
- **Strategy for solving the problem**
    - **Different depending on method we choose to pursue**
    - **For eval projects:**
        - What we plan to read
        - What we plan to test
        - What metrics we plan to use and why
        - If we are looking for some pattern, explain
        - If we believe some metrics do better for certain systems or others, explain
- **Collaboration plan**
    - **Who does what?**
        - same as roles listed in the beginning?
    - **How can we work simultaneously?**
        - research people can work together, while the writers can focus on reading the papers where we don't test code

***Note**:

Ask Tin how to translate a language model into sentiment analysis
Ask if this is a good data set, what should we be looking for in a data set
Division of work/roles

Data Set: https://ai.stanford.edu/~amaas/data/sentiment/

50,000 IMDb movie reviews annotated as positive/negative, including a training set

Tokenization Methods:

Splitting by space, Adjectives as tokens, Ignoring punctuation, TF-IDF weights, Stemming Included

Model: Support Vector Machines (SVM)

https://scikit-learn.org/stable/modules/svm.html

https://scikit-learn.org/stable/modules/linear_model.html

PROPOSAL:

Graham Harris, Judy Zheng, Christine Ma, Jong Heon Han
Data Set: https://ai.stanford.edu/~amaas/data/sentiment/

ROLES:

The team consists of four members. Graham Harris will be the writer, editor, and assistant researcher. Judy Zhang will lead the project research, and develop methods to tokenize the data inputted into the model. Christine Ma will implement the dataset for the computer to read, and also work on tokenization methods and general code maintenance. Jong Heon Han will be the data analyst and create an evaluation method to calculate the correlation between the answer key and test data.

PROBLEM STATEMENT:

There are many difficulties when attempting sentiment analysis of a given corpus because human language is so complex and has many different interpretations. Since movies reviews contain concise paragraphs of text with usually strong sentiment, they are a great starting point for analysis. Given a dataset of reviews with an assigned binary classified sentiment of positive and negative, represented in the data set as probabilistic range between 1 and 0 respectively, our team aims to test different tokenization methods on a reputable model in order to analyze the differences of how various tokens are weighted in their sentiment value. Several different tokenization methods will be compared to discern the best system.

EVALUATION PLAN:

Our team plans to use a variety of tokenization methods on a constant model to evaluate sentiment across the data set. On the NLP Progress website's page dedicated to sentiment analysis, there is a given list of accuracies for certain models on data sets pertaining to movie reviews. To keep the analysis consistent across trials, one model will be used, while various tokenizations of the input set will be developed as independent variables. We will then perform experiments by embedding these tokenizations into the model. The methods proposed are: splitting by spaces (control), adjectives as sole tokens, ignoring punctuation, TF-IDF weights, stemming included, and Support Vector Machines (SVM).

Since the dataset of 50,000 IMDb movie reviews has already been annotated as for sentiment (positive or negative), it will serve as the answer key and be compared to the output of the model. The model we decided on is a SVM model which as a regression, will output the sentiment analysis as a range, but for the sake of simplicity, the range will then be condensed into a grading of positive and negative where [0.0, 0.6) is considered negative and [0.6, 1.0] is considered positive. We will then compare and contrast the accuracy of the system to the annotated set to determine which system performs better on the bases of calculating recall, precision, and f-measure.

STRATEGY FOR SOLVING THE PROBLEM:

Each member will work on their assigned roles and communicate their findings. The team will start by reading various articles and papers on state-of-the-art sentiment analysis algorithms that solve the problem. By comparing the results of each tokenization though the model, we will draw conclusions regarding how sentiment has a role in language and how the development sentiment analysis by a computer aims to mimic understanding similar human evaluations.

COLLABORATION PLAN:

Our team will be able to collaborate by communicating our findings as we discover them to the rest of the team. As Graham and Judy have roles regarding research of the topic, they will communicate essential concepts and experiment recommendations for each tokenization process. Judy and Christine will work together on the initial stages of coding and implementing various tokenizations for the model. Then Jeong will evaluate the output and communicate the data to the rest of the group, as well as any figures and data representations. Then, everyone collectively will work together to draw conclusions as well as focus on improving and fine-tuning the experiment. While each member will be responsible for understanding the final report, Graham will synthesize the group's conclusions in a presentable and clear manner.

ARTICLES RELATED TO RESEARCH:

1.https://www.aclweb.org/anthology/W17-5410.pdf
Title: Breaking Sentiment Analysis of Movie Reviews
Author(s): By Ieva Staliūnaitė and Ben Bonfil (Utrecht University)
The article outlines how certain words and parts of speech have various meanings by context of the surrounding words. Essentially, the team concluded that since movie reviews allow for pragmatic and stylistic manipulations, it is difficult for systems to properly recognize sentiment. This research will be used to guide some of our tokenization methods, because our team will try to identify some difficulties and ambiguities that might be able to be improved upon.

2. https://arxiv.org/abs/1704.01444
Title: Learning to Generate Reviews and Discovering Sentiment
Author(s): Alec Radford, Rafal Jozefowicz, Ilya Sutskever
The work in this paper features sensitivity of learned representations of data models on distributions that they are trained on. Although our team's process will not be learning from our data set, we will be training it. This research team's sentiment analysis model was precise, interpretable, and manipulatable, which is a standard that our team will strive to achieve.

3. https://www.aclweb.org/anthology/Q17-1021.pdf
Title: Overcoming Language Variation in Sentiment Analysis with Social Attention
Author(s): Yi Yang and Jacob Eisenstein
This paper analyzes how different heterogeneous language variations make sentiment analysis more robust. The research team also used language variations from socially linked individuals to better perform sentiment analysis on Twitter users. This paper is of interest to our team language variation is a difficulty that POS-tagging and tokenization algorithms face. Our team can also pursue a tokenization method similar to theirs, where we group together reviews with similar language patterns.

4. https://arxiv.org/pdf/1905.05583.pdf
Title: How to Fine-Tune BERT for Text Classification?
Author(s): Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang
The goal of this research was to rigorously test different fine-tuning methods of BERT on text classification. Sentiment analysis tasks were run on the IMDb set (just like our research) and other sets,

for testing the BERT system. This paper also proposed other possible data sets for experimentation, such as Yelp food reviews.

5. https://arxiv.org/abs/1602.02373
Title: Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings
Author(s): Rie Johnson, Tong Zhang
This team used region embedding for text characterization. Their most interesting finding was that after combining two types of region embedding techniques, the best results were obtained on untrained data, which suggests that there were complementary text characterizations in their model. This is something our team will remain aware of - it may be beneficial to try to combine multiple tokenization methods for optimal results (if the methods are complementary).