Program Details

Welcome to BLAST AI's 2023 Summer Cohort! BLAST AI is dedicated to expanding access to AI research for high school students. This summer, you will learn the fundamentals of machine learning, work on a novel research project under the guidance of a mentor, and present your work at a camp-wide symposium.

This document contains information pertaining to the 2023 Summer Program. Please read through this full document before committing to attend the program and submitting your payment. Email info@blastai.org for more details.

Tuition

Tuition for the 8 week summer cohort will be $1160, inclusive of tax. Financial Aid forms were sent out to applicants who indicated financial need on their applications. If you discover that you can no longer attend, you will receive a refund if you inform BLAST more than 10 days in advance of the program's scheduled start date (see the Terms of Service for logistical details).

Program Dates

The camp will run from June 19 to August 13 (excluding weekends and holidays). The first two weeks will consist of the machine learning bootcamp (instruction is 8-12 PST), and the following six weeks will involve the completion of a guided research project. Feel free to review the Bootcamp Curriculum Schedule on Page 2.

Research Details

BLAST AI's Research program will run for 6 weeks. There will be workshops on the research process, office hours, and mentor sessions often. You will be matched to a research group of 5-6 people based on an interest form we will send you towards the end of the bootcamp. Please note that there is no guarantee that you will be matched into a group with your preferred interest, but we will try our best to accommodate groups based on preferences.

Terms of Service

By committing to BLAST AI, you accept the Terms of Service detailed in your acceptance letter.

Bootcamp Curriculum Schedule

Week 1

Day 1: Orientation and Python Crash Course

Orientation (45 min - 1 hr)

Environment and Notebook Set-Up (30 min)

Python Fundamentals Crash Course (2 hours)

Office Hours (1 hr)

Homework: Python Practice

Day 2: Data Manipulation

Python Homework Review (30 min)

Numpy Walkthrough (1 hour)

Pandas Walkthrough (2 hours)

Office Hours (1 hr)

Homework: Numpy and Pandas

Day 3: Data Visualization

Numpy and Pandas Homework Review (30 min)

Matplotlib Walkthrough (1 hour)

Seaborn Walkthrough (1.5 hour)

Office Hours (1 hr)

Homework: Matplotlib and Seaborn

Day 4: Intro to Machine Learning

Matplotlib and Seaborn Homework Review (30 min)

Intro to ML (1hr)

Linear Regression Lecture + Code Walkthrough (1 hour)

Logistic Regression Lecture + Code Walkthrough (1 hour)

Office Hours (1 hr)

Day 5: ML Basics and Ensembling

Capstone #1 Review: Data Science Essentials (30 min)

KNNs + Code Walkthrough (1 hour)

Decision Trees and Random Forest + Code Walkthrough (1 hour)

Ensembling Implementation (1 hour)

Week 2

Day 1: Neural Networks

Capstone #2 Review: ML Basics (1 hour)

Intro to Neural Networks + Math Theory (1.5 hr)

Neural Networks Continued + Code Walkthrough for Feed Forward NNs (1 hour)

Office Hours (1 hr)

Day 2: Computer Vision

Neural Networks Review (30 min)

Computer Vision CNN Lecture + Code Walkthrough (2 hours)

Office Hours (1 hr)

Day 3: Natural Language Processing

Capstone #2 Project Review (1.5 hours)

NLP Introduction + Code Walkthrough (1.5-2 hours)

Guest Lecture (1 hour)

Office Hours (1 hour)

Day 4: Transformers

Transformers Conceptually (1 hours)

Transformers Implementation (2 hour)

State of the art Transformers (1 hour)

Office Hours (1 hour)

Day 5: Reinforcement Learning and ChatGPT

Into to Q-Learning and PPO (1 hour)

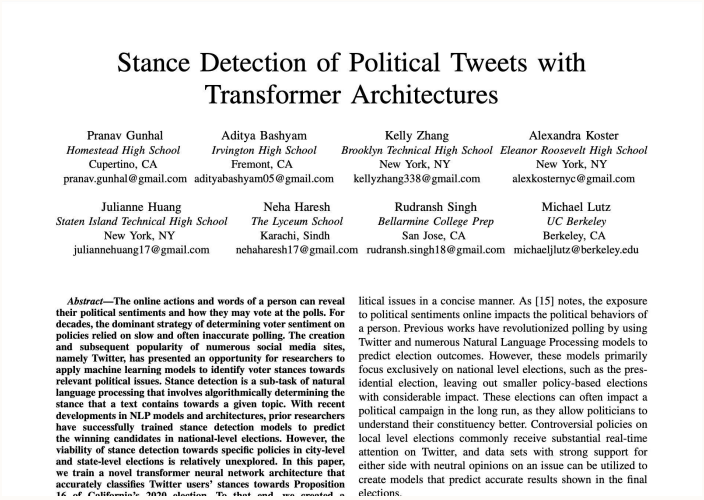
RL Mini-Project Implementation (2 hour)

RLHF and the InstructGPT framework of ChatGPT (30 mins)

Research

BLAST AI prides itself on the research that our students conduct. Guided by a mentor, every student completes a novel research project in a group of five additional students. Previous students have been accepted to PhD-Level conferences hosted by the Institute of Electrical and Electronics Engineers (IEEE) and the American Institute of Aeronautics and Astronautics (AIAA).

Sample research projects from our Summer 2022 Cohort are included below:

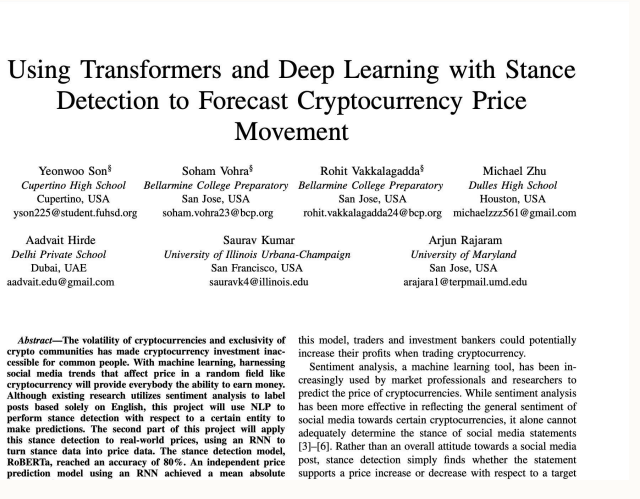


Stance Detection of Political Tweets with Transformer Architectures

Pranav Gunhal, *Homestead High School*, Cupertino, CA, pranav.gunhal@gmail.com
Aditya Bashyam, *Irvington High School*, Fremont, CA, adityabashyam05@gmail.com
Kelly Zhang, *Brooklyn Technical High School*, New York, NY, kellyzhang338@gmail.com
Alexandra Koster, *Eleanor Roosevelt High School*, New York, NY, alexkosternyc@gmail.com
Julianne Huang, *Staten Island Technical High School*, New York, NY, juliannehuang17@gmail.com
Neha Haresh, *The Lyceum School*, Karachi, Sindh, nehaharesh17@gmail.com
Rudransh Singh, *Bellarmine College Prep*, San Jose, CA, rudransh.singh18@gmail.com
Michael Lutz, *UC Berkeley*, Berkeley, CA, michaeljlutz@berkeley.edu

***Abstract*—The online actions and words of a person can reveal their political sentiments and how they may vote at the polls. For decades, the dominant strategy of determining voter sentiment on policies relied on slow and often inaccurate polling. The creation and subsequent popularity of numerous social media sites, namely Twitter, has presented an opportunity for researchers to apply machine learning models to identify voter stances towards relevant political issues. Stance detection is a sub-task of natural language processing that involves algorithmically determining the stance that a text contains towards a given topic. With recent developments in NLP models and architectures, prior researchers have successfully trained stance detection models to predict the winning candidates in national-level elections. However, the viability of stance detection towards specific policies in city-level and state-level elections is relatively unexplored. In this paper, we train a novel transformer neural network architecture that accurately classifies Twitter users' stances towards Proposition**

litical issues in a concise manner. As [15] notes, the exposure to political sentiments online impacts the political behaviors of a person. Previous works have revolutionized polling by using Twitter and numerous Natural Language Processing models to predict election outcomes. However, these models primarily focus exclusively on national level elections, such as the presidential election, leaving out smaller policy-based elections with considerable impact. These elections can often impact a political campaign in the long run, as they allow politicians to understand their constituency better. Controversial policies on local level elections commonly receive substantial real-time attention on Twitter, and data sets with strong support for either side with neutral opinions on an issue can be utilized to create models that predict accurate results shown in the final



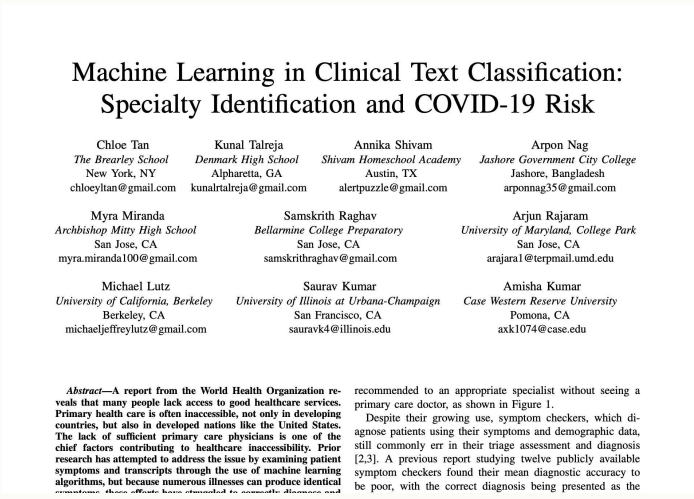
Using Transformers and Deep Learning with Stance Detection to Forecast Cryptocurrency Price Movement

Yeonwoo Son[§], *Cupertino High School*, Cupertino, USA, yson225@student.fuhsd.org
Soham Vohra[§], *Bellarmine College Preparatory*, San Jose, USA, soham.vohra23@bcp.org
Rohit Vakkalagadda[§], *Bellarmine College Preparatory*, San Jose, USA, rohit.vakkalagadda24@bcp.org
Michael Zhu, *Dulles High School*, Houston, USA, michaelzzz561@gmail.com
Aadvait Hirde, *Delhi Private School*, Dubai, UAE, aadvait.edu@gmail.com
Saurav Kumar, *University of Illinois Urbana-Champaign*, San Francisco, USA, sauravk4@illinois.edu
Arjun Rajaram, *University of Maryland*, San Jose, USA, arajara1@terpmail.umd.edu

***Abstract*—The volatility of cryptocurrencies and exclusivity of crypto communities has made cryptocurrency investment inaccessible for common people. With machine learning, harnessing social media trends that affect price in a random field like cryptocurrency will provide everybody the ability to earn money. Although existing research utilizes sentiment analysis to label posts based solely on English, this project will use NLP to perform stance detection with respect to a certain entity to make predictions. The second part of this project will apply this stance detection to real-world prices, using an RNN to turn stance data into price data. The stance detection model, RoBERTa, reached an accuracy of 80%. An independent price prediction model using an RNN achieved a mean absolute**

this model, traders and investment bankers could potentially increase their profits when trading cryptocurrency.

Sentiment analysis, a machine learning tool, has been increasingly used by market professionals and researchers to predict the price of cryptocurrencies. While sentiment analysis has been more effective in reflecting the general sentiment of social media towards certain cryptocurrencies, it alone cannot adequately determine the stance of social media statements [3]–[6]. Rather than an overall attitude towards a social media post, stance detection simply finds whether the statement supports a price increase or decrease with respect to a target



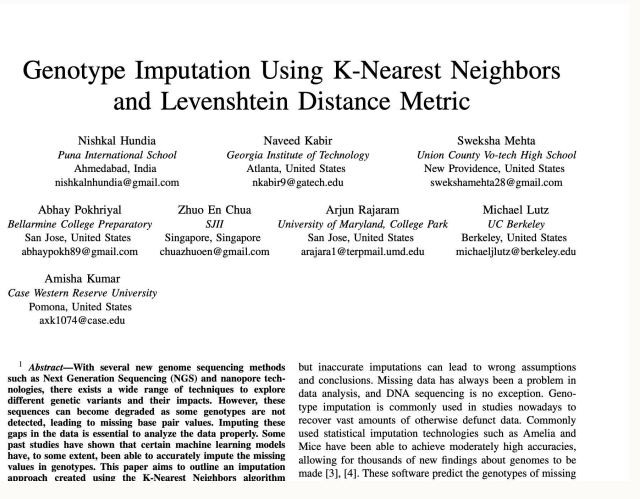
Machine Learning in Clinical Text Classification: Specialty Identification and COVID-19 Risk

Chloe Tan, *The Brearley School*, New York, NY, chloeyltan@gmail.com
Kunal Talreja, *Denmark High School*, Alpharetta, GA, kunalrtalreja@gmail.com
Annika Shivam, *Shivam Homeschool Academy*, Austin, TX, alertpuzzle@gmail.com
Arpon Nag, *Jashore Government City College*, Jashore, Bangladesh, arponnag35@gmail.com
Myra Miranda, *Archbishop Mitty High School*, San Jose, CA, myra.miranda100@gmail.com
Samskrith Raghav, *Bellarmine College Preparatory*, San Jose, CA, samskrithraghav@gmail.com
Arjun Rajaram, *University of Maryland, College Park*, San Jose, CA, arajara1@terpmail.umd.edu
Michael Lutz, *University of California, Berkeley*, Berkeley, CA, michaeljeffreylutz@gmail.com
Saurav Kumar, *University of Illinois at Urbana-Champaign*, San Francisco, CA, sauravk4@illinois.edu
Amisha Kumar, *Case Western Reserve University*, Pomona, CA, axk1074@case.edu

***Abstract*—A report from the World Health Organization reveals that many people lack access to good healthcare services. Primary health care is often inaccessible, not only in developing countries, but also in developed nations like the United States. The lack of sufficient primary care physicians is one of the chief factors contributing to healthcare inaccessibility. Prior research has attempted to address the issue by examining patient symptoms and transcripts through the use of machine learning algorithms, but because numerous illnesses can produce identical**

recommended to an appropriate specialist without seeing a primary care doctor, as shown in Figure 1.

Despite their growing use, symptom checkers, which diagnose patients using their symptoms and demographic data, still commonly err in their triage assessment and diagnosis [2,3]. A previous report studying twelve publicly available symptom checkers found their mean diagnostic accuracy to be poor, with the correct diagnosis being presented as the



Genotype Imputation Using K-Nearest Neighbors and Levenshtein Distance Metric

Nishkal Hundia, *Puna International School*, Ahmedabad, India, nishkalnhundia@gmail.com
Naveed Kabir, *Georgia Institute of Technology*, Atlanta, United States, nkabir9@gatech.edu
Sweksha Mehta, *Union County Vo-tech High School*, New Providence, United States, swekshamehta28@gmail.com
Abhay Pokhriyal, *Bellarmine College Preparatory*, San Jose, United States, abhaypokh89@gmail.com
Zhuo En Chua, *SJII*, Singapore, Singapore, chuazhuoen@gmail.com
Arjun Rajaram, *University of Maryland, College Park*, San Jose, United States, arajara1@terpmail.umd.edu
Michael Lutz, *UC Berkeley*, Berkeley, United States, michaeljlutz@berkeley.edu
Amisha Kumar, *Case Western Reserve University*, Pomona, United States, axk1074@case.edu

[1] ***Abstract*—With several new genome sequencing methods such as Next Generation Sequencing (NGS) and nanopore technologies, there exists a wide range of techniques to explore different genetic variants and their impacts. However, these sequences can become degraded as some genotypes are not detected, leading to missing base pair values. Imputing these gaps in the data is essential to analyze the data properly. Some past studies have shown that certain machine learning models have, to some extent, been able to accurately impute the missing values in genotypes. This paper aims to outline an imputation approach created using the K-Nearest Neighbors algorithm**

but inaccurate imputations can lead to wrong assumptions and conclusions. Missing data has always been a problem in data analysis, and DNA sequencing is no exception. Genotype imputation is commonly used in studies nowadays to recover vast amounts of otherwise defunct data. Commonly used statistical imputation technologies such as Amelia and Mice have been able to achieve moderately high accuracies, allowing for thousands of new findings about genomes to be made [3], [4]. These software predict the genotypes of missing

Feel free to view additional publications at: https://www.blastai.org/symposium/