Sentiment Analysis of India’s Tweets Regarding CPEC

by  
Jessica Crist

A capstone submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Data Analytics and Policy

Baltimore, Maryland December 2021

© 2021 Jessica Crist All Rights Reserved

Abstract

The Chinese Pakistan Economic Corridor (CPEC) is a potential economic stabilizer in the region; however, it encroaches upon India’s sovereignty and regional power dominance. Current literature regarding Indian public sentiment on CPEC is largely qualitative and focuses on themes such as military and economic strategy, and power dynamics. This research uses the quantitative approach of machine learning to analyze India’s sentiment on CPEC via Twitter postings. Tweets extracted from Twitter were prepared using Python’s Natural Language Toolkit (NLTK). Tweets were tokenized for feature extraction using Scikit-learn TfidfVectorizer, and analyzed using a logistic regression model that determined polarity probability with an 82.74% error rate, 84% precision rate, and 77% recall rate. Extracted features included terms, such as Gwadar, debt, investment, and Gilgit-Baltistan, aligned well with existing literature. The results demonstrate that this simple, easily replicated approach may serve well in near real-time analysis of public sentiment, even in locations where English is a non-primary language.

Table of Contents

Abstract………………...……………………………………………………………ii

Table of Content…………………………………………………………………......1

1. Introduction………………………………………………………………….2
2. Literature Review…………………………………………………………....3

2.1 Violation of Sovereignty…………………………………………………4

2.2 China’s Control of the Sea……………………………………………….6

2.3 Stability of Pakistan...................................................................................7

2.4 Regional Power and Containment……………………………………….9

1. Data and Methods…………………………………………………………...10

3.1 Data……………………………………………………………………...10

3.2 Methodology…………………………………………………………….12

1. Results……………………………………………………………………….14

4.1 Model Fit and Performance……………………………………………...14

4.2 Features…………………………………………......................................15

1. Conclusion………………………………………...........................................18

References....................................................................................................................19

Appendix A..................................................................................................................22

Appendix B..................................................................................................................24

**1. Introduction**

The China-Pakistan Economic Corridor (CPEC), a cornerstone of China’s Belt and Road Initiative (BRI), is the $62 billion dollar mobilization of the China’s 70-year diplomatic relationship with Pakistan.[[1]](#footnote-1) [[2]](#footnote-2) Through CPEC, China bestowed upon its “all-weather friend” Pakistan the financial credit necessary to modernize infrastructure, build out rail systems, stabilize the economy, and connect to Huawei fiber optics.[[3]](#footnote-3) [[4]](#footnote-4) Such an economic boon will almost certainly benefit the region, and previous sentiment analysis shows that in spite of some substantial setbacks, most Pakistani’s continue to favor the initiative.[[5]](#footnote-5) [[6]](#footnote-6) The project’s success has even inspired China to look to Taliban-governed Afghanistan for possible CPEC expansion. However, critical points of contention remain unresolved and could thwart CPEC’s timely completion. For example, the Indian government has lodged formal complaint for breach of sovereignty as the China-Indian Ocean connection will pass through Gilgit-Baltistan (a disputed territory currently occupied by Pakistan) and near Ladakh (a current flashpoint between China and India).[[7]](#footnote-7) [[8]](#footnote-8) [[9]](#footnote-9) The Indian government must now choose between stable diplomatic relations with a heavy-weight regional power (China) and territorial sovereignty. Additionally, in spite of historically tumultuous relations between India and Pakistan, stability in Pakistan is ultimately beneficial to India. Similarly, improved relations may help to dampen the nationalist ire that is so deeply rooted in Kashmir and Jammu. The progression of CPEC project completion ticks down the inevitable timeline for India to take action. Indian must weigh the loss of power and encroachment on sovereignty against the risk of creating a flashpoint for conflict, which has on numerous previous occasions escalated to the brink of war. The persistent threat of military conflict caused by regional hostilities over contested land and incursions against sovereignty intensify the need for all stakeholders and allies to vigilantly monitor local sentiment on politically sensitive topics such as CPEC.

In order to conduct analysis on India’s public sentiment on CPEC, this research uses logistic regression modeling and term frequency inverse document frequency on 1046 tweets originating from India and bordering regions. Features associated with the highest magnitudes of impact on polarity correspond with anticipated terminology based on CPEC-related qualitative literature. For example, key terms associated with negative sentiment include illegal, freebalochistan, gilgitbaltistan, corruption, loans, and debt. Conversely, key terms associated with positive sentiment included coastline, development investment, and connectivity. According to these findings. sentiment analysis of Tweets is a suitable means to help gauge public opinion on CPEC and may support additional sentiment analysis of politically sensitive topics in other countries even if English is not the primary or exclusive language for a population.

**2. Literature Review**

The Belt and Road Initiative (BRI) is China’s infrastructure development plan to connect China to global market in a manner that mirrors the historic Silk Road; it is heralded as China’s brand of globalization.[[10]](#footnote-10) In revitalizing historic trade routes, resource-dependent China seeks to secure access to resource-rich countries.[[11]](#footnote-11) The China-Pakistan Economic Corridor (CPEC), a cornerstone of China’s Belt and Road Initiative (BRI), is the $62 billion dollar mobilization of the China’s 70-year diplomatic relationship with Pakistan.[[12]](#footnote-12) [[13]](#footnote-13) China bestowed upon its “all-weather friend” the capital necessary to modernize infrastructure, build out rail systems, stabilize the economy, and connect to Huawei fiber optics.[[14]](#footnote-14) [[15]](#footnote-15) As part of BRI, CPEC will join China to Europe through South and Central Asian states via land routes and will create a marine corridor that gives China access to the Indian Ocean via deep-water port Gwadar.[[16]](#footnote-16) [[17]](#footnote-17)

The economic boon of CPEC will almost certainly benefit the larger region, and sentiment analysis shows that in spite of setbacks, most Pakistani’s continue to favor the initiative.[[18]](#footnote-18) [[19]](#footnote-19) CPEC’s projected success has inspired China to look to Afghanistan for future expansion. However, unresolved points of contention could yet hinder CPEC project completion in Pakistan, namely India’s absence from CPEC negotiations despite its position as a regional power. Distrustful of China’s military posturing, reluctant to subscribe to disadvantageous trade agreements, and traditionally opposed to development of disputed territories, the Indian government has voiced formal opposition to CPEC.[[20]](#footnote-20) [[21]](#footnote-21) Notwithstanding the opposition, CPEC progresses and each completed project diminishes the opportunity for dissent or action.

Although researchers have begun initial data analysis of Pakistanis’ sentiment towards CPEC,[[22]](#footnote-22) [[23]](#footnote-23) no such data analysis currently exists for India’s public sentiment on the issue. Similarly, literature on the topic of CPEC rarely focuses on India’s public sentiment in any way that is specific or measurable. Instead, the body of literature provides assessments of matters critical to India as a state, primarily focusing on four key themes: violation of India’s national sovereignty, China’s intentions for military use of Gwadar, implications of CPEC’s effect on stability in Pakistan, and erosion of India’s power and regional dominance.

**2.1 Violation of Sovereignty**

As part of CPEC, various sections of the 887-kilometer Karakoram Highway (KKH) that connect northern Pakistan with western China are under reconstructed and expanding.[[24]](#footnote-24) Improvement to the KKH is integral to reducing travel time between Islamabad and Gilgit (from 28 hours to 16 hours).The highway passes through Gilgit-Baltistan, a disputed region in northern Kashmir that is currently occupied by Pakistan. The Indian government has publicly rejected CPEC’s language that references Kashmir and lodged a formal complaint with the United Nations for breach of sovereignty.[[25]](#footnote-25) [[26]](#footnote-26) [[27]](#footnote-27) Some researchers suggest that CPEC may provide a useful vehicle for India and Pakistan to finally resolve their dispute of Kashmir, while others foresee CPEC’s development as providing de-facto legitimization to Pakistan’s rights to Gilgit-Baltistan.[[28]](#footnote-28) [[29]](#footnote-29) [[30]](#footnote-30) Additionally, CPEC roadways near Ladahk cause additional concerns for India’s sovereignty (Ladahk is a region in Kashmir that is bordered by Tibet to the east and Gilbit-Baltistan to the west).[[31]](#footnote-31) In 2020, China launched several incursions into India-controlled Ladakh, leading to deadly skirmishes that nearly escalated to war.[[32]](#footnote-32) China’s history of forceful development of disputed territory, as evident in the Doklam plateau crisis and Ladakh incursion, indicate China is unlikely to honor India’s diplomatic attempts to halt CPEC development in any of the disputed territories of Kashmir.[[33]](#footnote-33) [[34]](#footnote-34)

**2. 2 China’s Control of the Sea**

CPEC allocated 288 million USD to modernize deep-water Gwadar port, which became operational in 2016 and is now operationally controlled by China Overseas Port Holding Company (COPHC).[[35]](#footnote-35) The port gives China direct access to the Indian Ocean corridor, which is strategically important because it creates a 10,000-kilometer shortcut in the trade route between energy-rich countries such as Iran and energy-deficient China.[[36]](#footnote-36) [[37]](#footnote-37) [[38]](#footnote-38) Equally important, the new trade route bypasses the Malacca Straits, allowing China to avoid the significant vulnerabilities associate with transiting the straits, which are controlled by the United States and able to be blockaded in the event of military strife (known as the Malacca Dilemma).[[39]](#footnote-39)

India and the United States anticipate a Chinese Navy presence at Gwadar as indicated by the Pakistan-China joint military exercises in the Arabian Sea and China’s desire for expanded maritime prowess.[[40]](#footnote-40) [[41]](#footnote-41) [[42]](#footnote-42) A strong naval presence in Gwadar counters India’s ability to gain military dominance by immediately attacking Karachi during a conflict.[[43]](#footnote-43) Gwadar’s proximity to the Strait of Hormuz also provides an opportunity for Pakistan’s or China’s navy to monitor and intervene with maritime oil trade.[[44]](#footnote-44) [[45]](#footnote-45) India is attempting to counter developments in Gwadar by investing in Iran’s port of Chabahar.[[46]](#footnote-46) [[47]](#footnote-47) [[48]](#footnote-48) By creating a 220-kilometer road between Nimroz Province, Afghanistan and Chabahar, India is able to bypass a need for transshipment through Pakistan.[[49]](#footnote-49) However, investment in Chabahar will not necessarily deter aggression from China and Pakistan and in the event of conflict, China and Pakistan navies would likely disrupt the oil trade transiting west from the Middle East to Mumbai, Gujrat, and Kandla ports.[[50]](#footnote-50)

**2.3 Stability of Pakistan**

The magnitude of China’s CPEC investment will likely accomplish what aid provided by the United States could not - economic stability in Pakistan.[[51]](#footnote-51) An International Monetary Fund estimate anticipates that CPEC will add between 3 and 13 billion USD to Pakistan’s gross domestic product (GDP) through 2024, a calculation that is supported by early GPD increases in areas such as tourism in the cross-border regions.[[52]](#footnote-52)

CPEC projects are expected to create employment opportunities and stabilize economic growth, thereby reducing the drivers of terrorism in very poor areas such as Baluchistan Province.[[53]](#footnote-53) [[54]](#footnote-54) However, the Baloch National Movement, a militant group seeking autonomy for Baluchistan (home to Gwadar port), opposes CPEC and has redoubled efforts to thwart CPEC before China becomes further entrenched in the area.[[55]](#footnote-55) [[56]](#footnote-56) The Islamic State also poses a security threat, as made evident by the 2018 suicide attack in Mastung, Baluchistan Province.[[57]](#footnote-57) Similarly, Muslim separatist near the China-Pakistan border crossing may pose security threat for CPEC projects as routes connecting Uyghurs and the Federally Administered Tribal Areas would aid separatist mobilization.[[58]](#footnote-58) [[59]](#footnote-59) Pakistan’s perception of India’s role in backing Baloch separatists, adds the potential for political backlash against India in response to militant attacks on Chinese workers in Baluchistan or increased militant operations.[[60]](#footnote-60)  [[61]](#footnote-61) China’s and Pakistan’s leaders are committed to addressing security concerns cooperatively. China has funded routine patrol of the KKH and the Pakistani government deployed 9,000 Army soldiers and paramilitary forces as part of a Special Security Division protection force.[[62]](#footnote-62) [[63]](#footnote-63) [[64]](#footnote-64) The strategic importance of CPEC to China, as indicated by the staggering amount of money invested, suggests that China is unlikely to be deterred by security concerns.

Increased GDP may lead to a corresponding increase in Pakistan’s military spending, presumably aiding efforts to thwart militant organizations that pose a security threat. Conversely, an increase in Pakistan’s military spending could also initiate a military build-up and/or arms race with India, potentially disrupting the current status quo and decreasing regional stability.[[65]](#footnote-65) Strong economic and political alliances with China will reduce Pakistan’s need for cooperation with India. In fact, Pakistan may no longer feel pressure to attempt conciliation with India when it has China’s backing, a position that is easily exploited.[[66]](#footnote-66) [[67]](#footnote-67) For example, in the event that Pakistan is unable to make timely, complete loan repayments to Chinese banks, the Chinese government may exchange repayment extension or even loan forgiveness for Pakistan’s support in matters disadvantageous to India.

**2.4 Regional Power and Containment**

China’s approach to BRI predominantly focuses on encouraging diplomatic and economic partnerships through mutually beneficial business deals; however, the Chinese government is not above coercing those who are reluctant or in outright opposition.[[68]](#footnote-68) [[69]](#footnote-69) China’s BRI surrounds India in a clear strategy of confinement, forcing India to negotiate with China or remain excluded. Literature indicates that encirclement of India is deliberate and intended to undermine India’s power within the region, as India is China’s only hegemonic rival.[[70]](#footnote-70) In spite of India’s reluctance, China’s carefully constructed isolation of India by land and sea, and evidence to suggest that BRI will be successful without India’s inclusion, the Chinese government remains preferential to reaching a partnership with India.*[[71]](#footnote-71)* For example, during the 2008 Nuclear Supplies Group negotiation, China supported Indian exemptions as outlined in the US-India Civil Nuclear Agreement, demonstrating a willingness to court favor with India at the expense of Pakistan’s ire.[[72]](#footnote-72)

CPEC profoundly increases China’s presence and power throughout Central Asia. Long-term trade agreements, capital investments in infrastructure, and operational control of strategic entities such as ports lock in partnerships and partner states’ dependency on China. Through arrangements like CPEC, states exchange strategic advantage for anticipated future prosperity. CPEC, being the flagship of BRI, offers a significant amount of strategic advantage because of the Pakistan-based terrestrial and maritime access, but also as a prize puzzle piece that neatly connects with other regional partnerships with China. The BRI framework encompasses 68 countries and up to 40 percent of the global GDP.[[73]](#footnote-73) [[74]](#footnote-74)Most researchers anticipate that CPEC will benefit not only Pakistan, but the surrounding region; however, benefits specific to India, beyond those of regional stability, have not yet been demonstrated.[[75]](#footnote-75) Instead, India must carefully mitigate loss of power and influence to economic-powerhouse China.[[76]](#footnote-76)

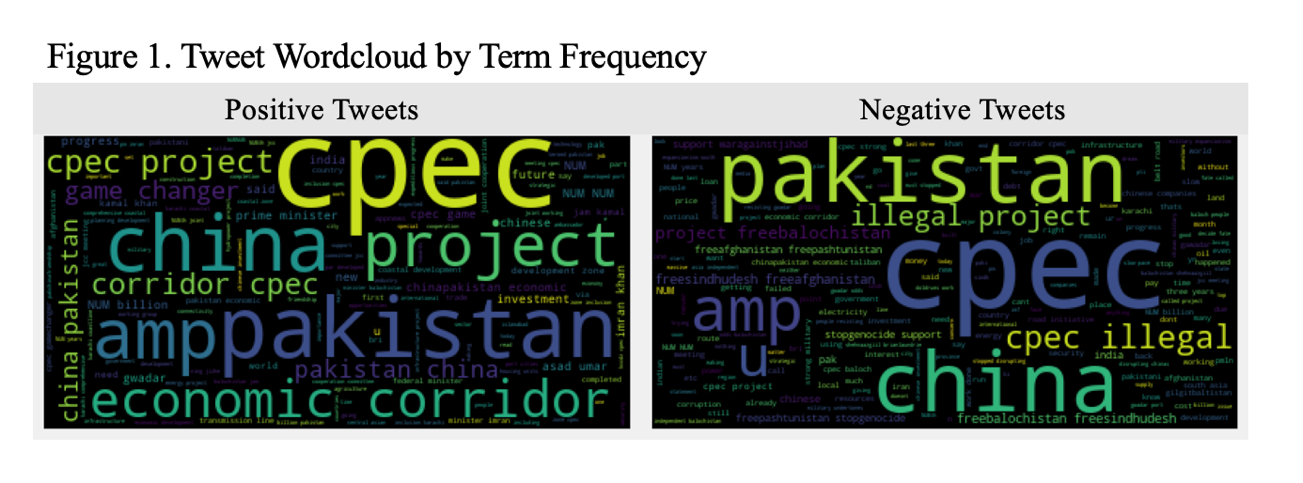
**3. Data and Methods**

**3.1 Data**

This research uses a repository of tweets extracted using Python’s tweepy library to access Twitter’s developer application programming interface (API). The script targeted tweets containing CPEC-related key words and tweeted after October 1, 2013 (the inception of CPEC).[[77]](#footnote-77) The script also targeted tweets within 3,000 km radius of Nagpur, India in an effort to limit tweet origins to India and bordering areas.[[78]](#footnote-78) Because retweeting is a common form of expressing agreement or desire to promote a sentiment, retweets were included in the collection.

The extracted tweets were converted from a .csv to a data frame in which each tweet is a row. In preparation for analysis, using Python’s Natural Language Toolkit (NLTK), tweets were duplicated and the data was cleaned by changing tweets to all lower case, and removing punctuation, emoji, English stop words, and special characters. Although the use of emoji has been found useful in specialized research, their incorporation is beyond the scope of this capstone. From the collected pool of 23,115 tweets, 2,026 post-processing tweets remained for analysis. Of the 2,026 tweets available, 1,000 were manually labeled for training the machine model. Amongst the tweets labeled as positive, the most frequency terms included: cpec, pakistan, china, project, and economic corridor. Amongst the tweets labeled as negative, the most frequent terms included: pakistan, illegal, project, cpec, and china (see Figure 1. Tweet Wordcloud by Term Frequency for a depiction of the Tweet term frequency.)

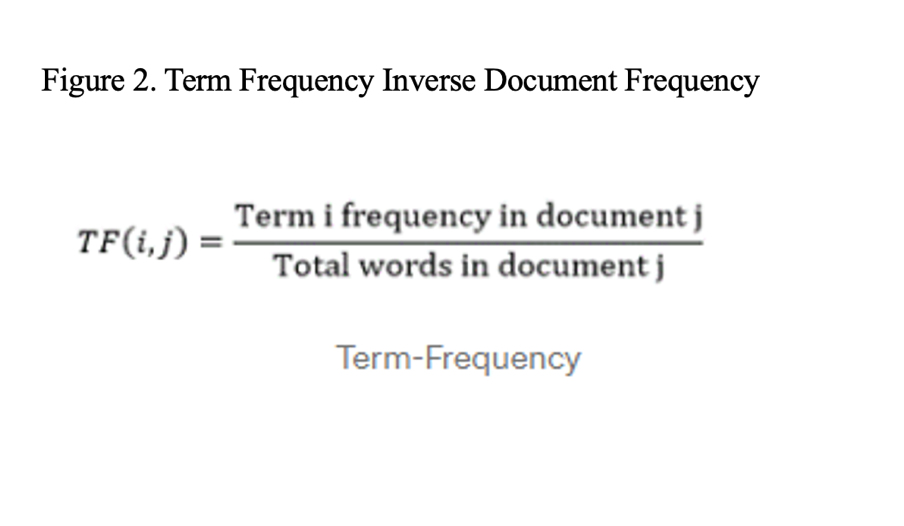
The limitations of this data source stem from the inherent limitations associated with any social media platform in that the population is self-selecting. Additionally, the public nature of posts may influence users’ opinion portrayal or self-branding in an effort to curry favor with an intended audience. Using a Twitter developer’s account also includes built-in constraints such as maximum endpoint request rates and limits on geolocational targeting. Additionally, focusing on tweets written predominantly in the English language, reduces the sample to tweets written by individuals who are willing and able to express their opinion in English. Since English is one of the two official languages in India, the associated risk of sampling bias is unlikely to undermine research findings.



**3.2 Methodology**

This research uses Term Frequency and Inverse Document Frequency (TFIDF) and logistic regression to conduct sentiment analysis. Sentiment analysis refers to “the use of natural language processing, text analysis, computational linguistic, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.”[[79]](#footnote-79)

TFIDF was used for feature engineering and extraction.[[80]](#footnote-80) Term frequency is the total count of the unique words contained within a tweet, and inverse document frequency is used to weigh the importance or rarity of the words (see Figure 2. Term Frequency Inverse Documentation Frequency for a simplified annotation of TFIDF formula). In text analysis and natural language processing, TFIDF provides a statistical measure to evaluate term relevance within a larger collection, from which features can be extracted. The goal is to identify terms that are important to the individual tweets within the corpus of tweets.[[81]](#footnote-81)



Logistic regression was used as the classifier by which TFIDF was deployed. Supervised machine learning was selected over unsupervised techniques as supervised machine learning has historically outperformed unsupervised methods in sentiment analysis. [[82]](#footnote-82) Although advancements in semantic resources and hybridization of supervised and unsupervised learning have reduced error rates for unsupervised learning, on the whole both methods remain similar in performance when analytic rigor is maintained.[[83]](#footnote-83) Amongst supervised learning techniques, although no one technique has typically outperformed the other in application across all domains,[[84]](#footnote-84) logistic regression was chosen based on its favorabe feature interpretability.[[85]](#footnote-85) [[86]](#footnote-86) Thus the results were more likely to be readily usable in this research and easily replicable for similar policy-related endeavors.

**4. Results**

**4.1 Model Fit and Performance**

Using logistic regression machine modeling for sentiment analysis of India-origin Tweets on the topic of CPEC successfully demonstrates the ability to classify sentiment with an accuracy rate of 82.74%. Additionally, feature extraction provides meaningful insight on key positive and negative terms and their magnitude as designated by model coefficients. The model was most accurate in predicting positive tweets.

In this research of the 2045 unique tweets, 1,000 tweets are manually labeled for use in model training. Among the training tweets, 470 are positive (1) and 530 are negative (0). Using the remaining 1045 unlabeled testing tweets, a logistic regression model is able to classify (based in probability of sentiment polarity) with an accuracy rate of 82.74% (see Table 1. Logistic Regression Model Confusion Matrix). The recall rate for positive tweets (96%) is significantly higher than for negative tweets (58%), indicating that the model struggled most to predict negative tweets as negative. The F1-score, which measures the model’s accuracy on this dataset by combing precision and recall, indicates that the model is relatively accurate, in spite of the limitations in classifying negative tweets as negative (see Table 2. Logistic Regression Model Classification Report).

A C value of 1 is assigned to correct for initial model overfit at values of 0.1 and lower. Regularization L1 penalty (Lasso Regression) outperformed L2 regularization (Ridge Regression).[[87]](#footnote-87) This is consistent with expectations as Lasso shrinks the coefficient of less important features, so it is generally better suited for feature selection in cases with a large selection of features such as those in this study. Based on the size of the data set and incorporation of an L1 penalty, the saga solver parameter is selected in order to optimize the objective function.[[88]](#footnote-88)

**Table 1:** Logistic Regression Model Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 307 | 223 |
| Actual Positive | 36 | 935 |

**Table 2:** Logistic Regression Model Classification Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 0.90 | 0.58 | 0.70 | 530 |
| Positive | 0.81 | 0.96 | 0.88 | 971 |
|  | | | | |
| Accuracy |  |  | 0.83 | 1501 |
| Macro Average | 0.84 | 0.77 | 0.79 | 1501 |
| Weighted Average | 0.85 | 0.83 | 0.82 | 1501 |

**4.2 Features**

Extracted features are found to be consistent with expected terms of importance based on common themes identified in the Literature Review. The model identifies 4973 features of which 91 have an absolute coefficient values of greater than 0.000000001 (1e9) and are extracted as having a value of interest. From the 91 extracted features, Table 3 includes a subset of the top relevant positive and negative features (for a list of all features, see Appendix 2).

The top positive features include words such as industry, friendship, coastline, cooperation, investment, changer, pm, and minister, which thematically match with expectations. Within the region believe that China’s investment of Pakistan is touted as a game-changing opportunity to develop strategic industries and areas such as Karachi. When discussing the positive aspects of CPEC, Tweets tend to include terms such as investment and future.

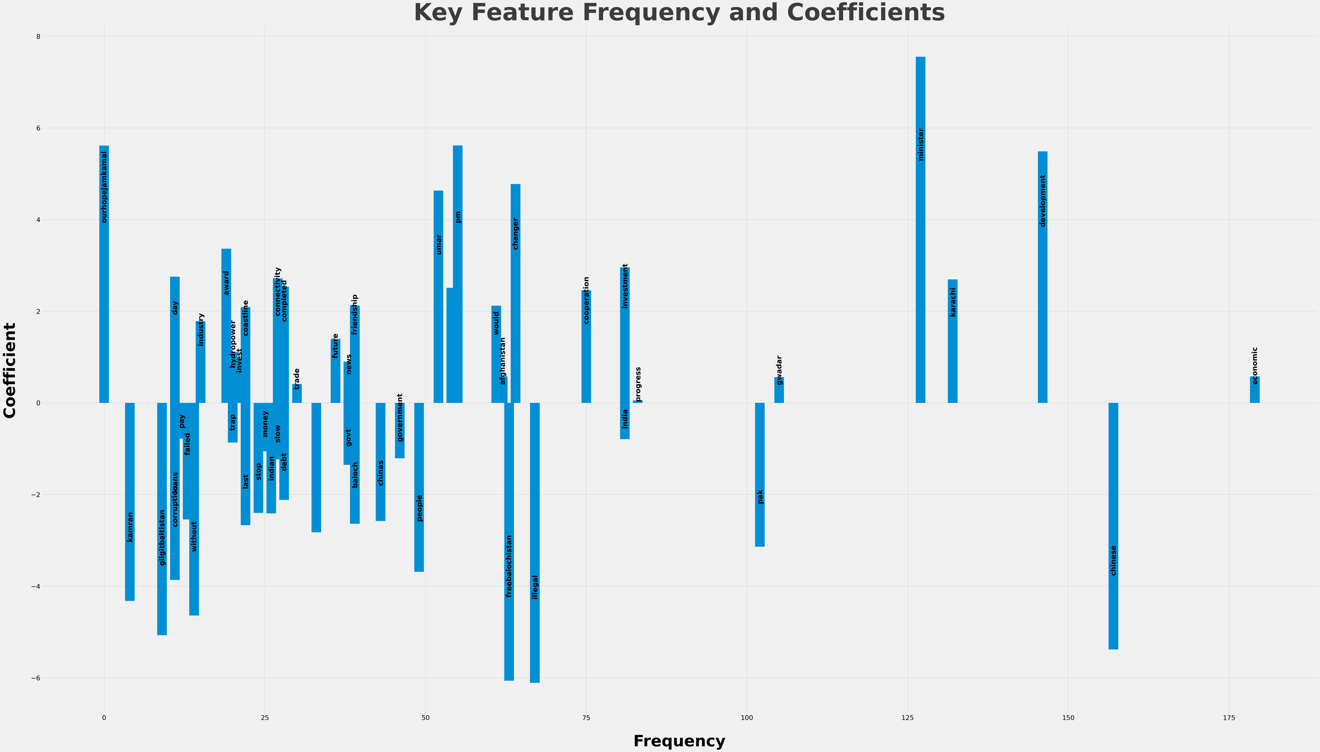
The top negative features include illegal, freebalochistan, gilgitbaltistan, corruption, loans, chinese, and debt. These features are in keeping with the sentiment that China’s high-interest rate loans will create a debt trap for Pakistan and that the CPEC infrastructure is being illegally built on contested lands such as Gilgit Baltistan and the autonomous region of Baluchistan. Negative Tweets more commonly include terms highlighting the failures or slow progress of CPEC projects and associated losses such as money, land, and autonomy (i.e., becoming a Chinese “colony”).

As expected, some terms appear frequently within the Tweets, such as Pakistan, Chinese, and development. Other terms, such as freebalochistan suggest a commonly used hashtag for group themes or movements. Variations in word spelling appear to be relevant to sentiment polarity; for example, reference to Pakistan as a proper noun is associated with positive coefficients, whereas Paks as a more colloquial or informal term is associated with a negative coefficient. Term frequency was likely influence by the significance of a word to the CPEC topic, as indicated by the frequency of Gwadar or its alternative spelling Gawadar. Other terms such as Afghanistan and Karachi are likely linked to spikes in CPEC-related Tweets on short term trending topics. For example, the official decision to include Karachi in CPEC and China’s contemplation of expanding CPEC to include Afghanistan following the Taliban’s rise to power likely accounts for the associated higher term frequencies. Additionally, the Standard Twitter Developer’s Account is bias towards recently posted Tweets and favors them during API extraction unless additional permissions are established to allow for historical pulls. As such, using this methodology for feature extraction is likely to benefit current trends in term significance rather than broad-reaching comparison. See Table 3. Key Feature Frequency and Coefficients for a visual representation of the frequency of extracted features with most apparent relevance.

**Table 3.** Coefficients and Frequency of Test Tweets Prominent Positive and Negative Features\*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Positive  Feature | Coefficient | Frequency | Negative Feature | Coefficient | Frequency |
| progress | 0.076638239 | 32 | illegal | -6.109758258 | 52 |
| gwadar | 0.561304551 | 84 | freebalochistan | -6.064753674 | 52 |
| afghanistan | 0.574098814 | 63 | chinese | -5.381151209 | 141 |
| invest | 0.955098148 | 10 | gilgitbaltistan | -5.069178471 | 8 |
| hydropower | 1.105958358 | 1 | corruption | -3.865338668 | 10 |
| coastline | 2.088405388 | 8 | people | -3.685732932 | 28 |
| friendship | 2.124580272 | 11 | pak | -3.139310453 | 53 |
| pakistan | 2.373098746 | 303 | loans | -2.797816394 | 17 |
| cooperation | 2.45659374 | 34 | baloch | -2.638140963 | 37 |
| pakistans | 2.511555584 | 33 | chinas | -2.578135651 | 14 |
| completed | 2.534580218 | 16 | indian | -2.412181116 | 23 |
| karachi | 2.694802113 | 61 | debt | -2.117759369 | 13 |
| connectivity | 2.716950925 | 6 | ruined | -2.073174608 | 3 |
| investment | 2.956744196 | 10 | failed | -1.630304016 | 13 |
| award | 3.365365415 | 3 | slow | -1.233511367 | 7 |
| changer | 4.777076669 | 13 | money | -1.056448337 | 13 |
| development | 5.488686728 | 53 | trap | -0.865926371 | 15 |
| ourhopejamkamal | 5.615909175 | 2 | india | -0.790293224 | 58 |
| pm | 5.616400354 | 23 | colony | -0.528884431 | 3 |
| minister | 7.554776195 | 63 | gawadar | -0.373752767 | 4 |
| \* Feature terms lacking in relevance such as “use”, “ur”, and “held” were excluded from this table | | | | | |

Figure 3. Key Feature Frequency and Coefficients



\* The term Pakistan (frequency 507) was removed from this graph as its status as an outlier is expected, given Pakistan’s relation to CPEC and its inclusion would not have enhanced the understandability of the other features

**5. Conclusion**

The possible applications of sentiment analysis are vast; however, in spite of the successful sentiment analysis of this research, more advanced analysis on the topic is warranted. For example, understanding divergence in public sentiment by region, religion, or ethnicity would illuminate disparity in access and use of definitive information, susceptibility to local propaganda or cultural influence, and relationships with high magnitude features (e.g., populations living in closer geographic proximity to militarily contested areas may more or less positive of China’s political reach into nearby states).

The successful sentiment analysis of India public sentiment towards CPEC indicates a utility in applying similar methodology against other politically sensitive topics with other populations in which English is not a primary or exclusive language. Additionally, as lexicons and natural language toolkits become more robust in non-English languages or hybrid languages such as Hinglish (Hindu and English), likely machine learning models will be able to identify polarity probability with increased levels of accuracy, further expanding the suitability of sentiment analysis in policy related analytics.

This research leveraged logistic regression and TFIDF, a relatively simple approach to sentiment analysis that lends well to easy replication and yields readily interpretable results. In using a methodology such as this, governments may be able to monitor public sentiment in near real-time for topics of immerging importance or crises as they develop. Governments that use sentiment analysis as a tool to help inform decision making, would likely enhance their ability to respond with improved precision and relevance for matters such as calming civil unrest, countering conspiracy with increased transparency, or adapting policies to match shifts in societal expectations of leadership. In using sentiment analysis, governments have the capacity to understand their constituent’s concerns and preferences in a non-invasive manner that complements other, more traditional social sciences research.

References

Amina, Bibi and Tayyaba Azim. “SCANCPECLENS: A Framework for Automatic Lexicon Generation and Sentiment Analysis of Micro Blogging Data on China Pakistan Economic Corridor.” IEEE (2019): 133876 - 133887. Doi 10.1109/ACCESS.2019.2940528.

Awasthi, Soumya. “China-led BRI and its Impact on India with Specific Reference to CPEC.” Journal of International Affairs 2 (2018), accessed September 1, 202, https://doi.org/10.3126/joia.v2i1.22592

Cai, Peter. “Why India Distrusts China’s One Belt One Road Initiative.” The Lowly Institute, (2016). Https://www.lowyinstitute.org/the-interpreter/why-india-distrusts-chinas-one-belt-one-road-initiative

Chen, Xiangming, S.K. Joseph, and Hamna Tariq. "Betting Big on CPEC." *European Financial Review* (2018); 61-70. http://www.europeanfinancialreview.com/?p=21658

Chinoy, Sujan. “CPEC: China’s designs, Pakistan’s Ambivalence and India’s Opposition.” *Hindustan Times* (2021). https://www.hindustantimes.com/opinion/cpec-china-s-designs-pakistan-s-ambivalence-and-india-s-opposition-101611066633406

DD News. “China Ignores India’s Sovereignty Concerns, Plans to extend Controversial CPEC to Afghanistan.” *DD News* (2021). <https://ddnews.gov.in/international/china-ignores-india%E2%80%99s-sovereignty-concerns-plans-extend-controversial-cpec-afghanistan>

Houreld, Katharine, “China and Pakistan Launch Economic Corridor Plan Worth $45 Billion.” *Reuters* (2015). Https://www.reuters.com/article/us-pakistan-china-idUSKBN0NA12T20150420

Hussain, Fakhar, Dr. MianSaeed Ahmad, Saadat Nawaz, Syed AqeelHaider, and Muhammad Atif. “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective.” *EEO* 20 (2021): 3764-3770. Doi:10.17051/ilkonline.2021.05.414.

Khan, Hafeez Ullah. “Regional Security Threat to CPEC: A Strategic Overview.” *Journal of the Research Society of Pakistan* 56 (2019). https://www.researchgate.net/publication/335095537\_Regional\_Security\_Threats\_to\_CPEC\_A\_Strategic\_Overview

Kuhn, Max and Johnson, Kjell. Feature Engineering and Selection: A Practical Approach for Predictive Models. Milton: CRC Press LLC (2019): 114-120.

Latif, Muhammad Ijaz and Muhammad Tayyab Zia. “Strategic Dimensions of CPEC: Role of Regional and International Powers.” *Journal of Business and Social Review in Emerging Economies*, 6 (2020):1561-1569. Https://doi.org/10.26710/jbsee.v6i4.1507

Mamgain, Nehal, Ekta Mehta, Ankush Mittal, and Gaurav Bhatt. "Sentiment analysis of top colleges in India using Twitter data." International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT) (2016):525-530. Doi:10.1109/ICCTICT.2016.7514636.

Mashalkar, Atharva. “Sentiment Analysis using Logistic Regression and Naïve Bayes.” Towards Data Science (2020). Https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b

Nagpal, Anuja. “L1 and L2 Regularization Methods.” *Towards Data Science* (2017). Https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

Neri, Federico, Carlo Aliprandi, Federico Capeci, Montse Cuadros and Tomas By. “Sentiment Analysis on Social Media.” International Conference on Advances in Social Networks Analysis and Mining (2012). Doi:10.1109/ASONAM.2012.164.

Palmer, Alex W. Palmer. “The Man Behind China’s Aggressive New Voice.” *The New York Times Magazine* (2021). https://www.nytimes.com/2021/07/07/magazine/china-diplomacy-twitter-zhao-lijian.html

Pothabattula, Santosh Kumar. “The Math and Intuition behind Logistic Regression.” *Analytics Vidhya* (2019). Https://medium.com/analytics-vidhya/a-complete-understanding-of-how-the-logistic-regression-can-perform-classification-a8e951d31c76

Qianqian, Li and Liu Yijun. “The China-Pakistan Economic Corridor: The Pakistani Media Attitudes Perspective.” *Technology in Society* 62 (2020): 2-5. <https://doi.org/10.1016/j.techsoc.2020.101303>

Ratna, Paras. “India’s CPEC Dilemma – To Participate or not to Participate.” East Asia Research Programme. http://earp.in/en/indias-cpec-dilemma-to-be-or-not-to-be/

Sacks, David. “The China-Pakistan Economic Corridor-Hard Reality Greets BRI’s Signature Initiative.” *Council on Foreign Relations* (2021). https://www.cfr.org/blog/china-pakistan-economic-corridor-hard-reality-greets-bris-signature-initiative.

Tarapore, Arzan. “The Crisis After the Crisis: How Ladakh will Shape India’s Competition with China.” Lowy Institute (2021). Https://www.lowyinstitute.org/publications/crisis-after-crisis-how-ladakh-will-shape-india-s-competition-china

Wagner, Christian. “The effects of the China-Pakistan Economic Corridor on India-Pakistan relations.” (SWP Comment, 25/2016). Berlin: Stiftung Wissenschaft und Politik -SWP- Deutsches Institut für Internationale Politik und Sicherheit. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-46898-4

APPENDIX A: Features with Coefficients Greater than 1e9

|  |
| --- |
|  |
| illegal - -6.109758257530541 |
| freebalochistan - -6.064753673902656 |
| chinese - -5.381151209186677 |
| gilgitbaltistan - -5.06917847092516 |
| without - -4.640519027548493 |
| kamran - -4.323126025482625 |
| corruption - -3.8653386684623383 |
| people - -3.685732932422443 |
| pak - -3.1393104527364044 |
| use - -2.8880152280978693 |
| one - -2.8239614365786636 |
| loans - -2.797816394117057 |
| last - -2.668973884012108 |
| baloch - -2.63814096348394 |
| chinas - -2.57813565129067 |
| three - -2.5439001732375845 |
| indian - -2.4121811157335347 |
| stop - -2.4002930273091883 |
| cant - -2.3599478031362118 |
| debt - -2.1177593686396956 |
| ruined - -2.073174608040544 |
| losing - -2.072546845721353 |
| anything - -2.0031159578651327 |
| even - -1.9013368421641341 |
| failed - -1.630304016400995 |
| ur - -1.4773389173250902 |
| govt - -1.3536775577704536 |
| called - -1.2451774218762486 |
| slow - -1.2335113675121296 |
| government - -1.2090482474317552 |
| isnt - -1.171619905021707 |
| money - -1.0564483370755382 |
| never - -0.9423845012776071 |
| pakistanis - -0.8831697559472377 |
| trap - -0.8659263713174588 |
| india - -0.7902932244991939 |
| pay - -0.7788909726644541 |
| years - -0.7232059988684957 |
| colony - -0.5288844319690583 |
| thats - -0.45571590868628137 |
| gawadar - -0.37375276737117125 |
| china - -0.2941124134890186 |
| land - -0.27003611635058955 |
| trying - -0.24850900247030047 |
| military - -0.17610909608428071 |
| electricity - -0.15497589257630526 |
| amp - -0.0075932192385067925 |
| says - 0.005240715845733706 |
| read - 0.038344740321184276 |
| progress - 0.049636821977268625 |
| regional - 0.07663823867808482 |
| moving - 0.09687831092044841 |
| pakchinamediaforum - 0.14576113418402017 |
| ecommerce - 0.1990275594381753 |
| transmission - 0.2906491929176627 |
| appnews - 0.3623325125920141 |
| help - 0.3981665141930121 |
| trade - 0.4109648502729953 |
| welcome - 0.43720291848524967 |
| speed - 0.5435892594736653 |
| gwadar - 0.5613045506547827 |
| afghanistan - 0.5740988139440507 |
| economic - 0.5774108619879242 |
| news - 0.898252478467543 |
| invest - 0.955098147680642 |
| chinapakistan - 1.0006859101188281 |
| completion - 1.0193341248892658 |
| hydropower - 1.1059583579296781 |
| pakchina - 1.2342061751468885 |
| best - 1.3519677602263251 |
| future - 1.3954123810285641 |
| numth - 1.4632590238892988 |
| industry - 1.7821409834547417 |
| coastline - 2.0884053882948614 |
| would - 2.120363259183783 |
| friendship - 2.1245802716599935 |
| pakistan - 2.373098746129387 |
| cooperation - 2.4565937400594935 |
| pakistans - 2.511555583502989 |
| completed - 2.534580217629335 |
| held - 2.672092233788312 |
| karachi - 2.6948021134726154 |
| connectivity - 2.7169509247680765 |
| day - 2.75346299158805 |
| investment - 2.9567441964298853 |
| award - 3.3653654149835677 |
| umar - 4.632169295658351 |
| changer - 4.777076669185278 |
| development - 5.488686727704367 |
| ourhopejamkamal - 5.615909175337378 |
| pm - 5.6164003544027645 |
| minister - 7.554776195450144 |

APPENDIX B: Python Script

Extracting Tweets from Twitter API

import tweepy # pip install tweepy / brew install tweepy (if on a mac... probably)

#import pandas # pip install pandas / brew install pandas (if on a mac... probably)

import re

# set up and read in keys

CONSUMER\_KEY = ''

CONSUMER\_SECRET = ''

ACCESS\_TOKEN = ''

ACCESS\_TOKEN\_SECRET = ''

try:

with open("keys.txt","r") as f:

keys = f.readlines()

CONSUMER\_KEY = keys[0].rstrip()

CONSUMER\_SECRET = keys[1].rstrip()

ACCESS\_TOKEN = keys[2].rstrip()

ACCESS\_TOKEN\_SECRET = keys[3].rstrip()

except IOError:

print("make sure that keys.txt exists in the same directory as this file.")

# authenticate

auth = tweepy.OAuthHandler(CONSUMER\_KEY, CONSUMER\_SECRET)

auth.set\_access\_token(ACCESS\_TOKEN, ACCESS\_TOKEN\_SECRET)

api = tweepy.API(auth, wait\_on\_rate\_limit=True)

# get the tweets. returns a list of JSON formatted strings. I think...

def get\_tweets(keyword="cpec", date\_since="2013-01-01", max\_tweets="100"):

tweets = tweepy.Cursor(api.search,

keyword, #contains the keyword

geocode="20.5937,78.9629,1000km", #from India

lang="en", #english

since=date\_since, #tweeted after date

tweet\_mode="extended" #gets the full tweet. If this isnt set, change tweet.full\_text to tweet.text.

).items(max\_tweets) #only get max\_tweets number of tweets. Omit for unlimited tweets.

return tweets

# takes a list of JSON formatted strings and returns a list of just the tweet text

def strip\_tweets(tweets):

tweet\_list = []

# only get the text of the tweet and store in tweet\_list

i=1

for tweet in tweets:

tweet\_list.append(tweet.full\_text)

# log

print("gotten", i, "tweets so far!")

#print(i, tweet.full\_text)

i+=1

return tweet\_list

# fix up some of the punctuation and whitespaces to make storage easier/csv format possible.

# tweets definitely need a lot more cleaning but this shouldn't hurt the data too badly.

def fix\_tweets(tweet\_list):

fixed\_tweets = []

for tweet in tweet\_list:

s = tweet

s = re.sub(r'^https?:\/\/.\*[\r\n]\*', '', s, flags=re.MULTILINE) # removes links! (yoinked from stack overflow)

s = s.replace('\n', ' ').replace(',', ' ').replace('.',' ') #ew. (replaced punctation with spaces to avoid weird conglomerations of words)

fixed\_tweets.append(s)

return fixed\_tweets

# store the tweets into a CSV file named "tweets.csv" that can be read / analyzed later on!

# currently APPENDS to the .csv file. this can be easily changed to write by modifying the open() parameters!

def store\_tweets(tweet\_list:list):

with open('tweets2.csv', 'a', encoding="utf-8") as file: #need to specify utf-8 encoding for some reason. Dont ask me why.

for tweet\_text in tweet\_list:

#print(tweet\_text)

file.write(tweet\_text+",")

file.close()

# get, strip, clean, and store!

def main():

tweets = get\_tweets("cpec", "2013-01-01", 100)

tweet\_list = strip\_tweets(tweets)

tweet\_list = fix\_tweets(tweet\_list)

store\_tweets(tweet\_list)

if \_\_name\_\_ == "\_\_main\_\_":

main()

Cleaning, Processing, Tokenizing, and Modeling

from nltk import text

import pandas as pd

import numpy as np

from wordcloud import WordCloud

import matplotlib.pylab as plt

import matplotlib as mpl

import seaborn as sns

import re

import ssl

import nltk

from nltk.corpus.reader.knbc import test

from nltk.tokenize.regexp import WhitespaceTokenizer

ssl.\_create\_default\_https\_context = ssl.\_create\_stdlib\_context

from nltk import word\_tokenize, pos\_tag

from nltk.corpus import stopwords

from nltk.corpus.reader import CorpusReader

from nltk.internals import deprecated

from nltk.probability import FreqDist

from nltk.tokenize import word\_tokenize

from nltk.tokenize import TweetTokenizer

from nltk.stem import SnowballStemmer

from nltk.corpus import wordnet

from nltk.stem.wordnet import WordNetLemmatizer

#nltk.download('stopwords')

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.svm import LinearSVC

from sklearn.pipeline import Pipeline

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_selection import SelectKBest, chi2

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import classification\_report

import neattext.functions as nfx

import random

from sklearn.metrics import accuracy\_score

from sklearn.metrics import confusion\_matrix

from collections import Counter

from collections import defaultdict

test\_tweets = pd.read\_csv('./testing\_tweets\_1.csv', delimiter=',')

train\_tweets = pd.read\_csv('./training\_tweets\_1.csv', delimiter=',', encoding= 'unicode\_escape')

#print(train\_tweets)

#make all lowercase

test\_tweets["Cleaned\_Tweet"] = test\_tweets["Tweet"].str.lower()

train\_tweets["Cleaned\_Tweet"] = train\_tweets["Tweet"].str.lower()

#print(test\_tweets.head())

#print(train\_tweets.head())

#remove numbers and special characters

p = re.compile (r"\d+")

test\_tweets["Cleaned\_Tweet"] = [p.sub ('NUM', x) for x in test\_tweets["Cleaned\_Tweet"]]

p2=re.compile("@[A-Za-z0-9\_]+")

test\_tweets["Cleaned\_Tweet"] = [p2.sub ('', x) for x in test\_tweets["Cleaned\_Tweet"]]

test\_tweets["Cleaned\_Tweet"] = test\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_emojis)

test\_tweets["Cleaned\_Tweet"] = test\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_multiple\_spaces)

test\_tweets["Cleaned\_Tweet"] = test\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_urls)

test\_tweets["Cleaned\_Tweet"] = test\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_punctuations)

test\_tweets["Cleaned\_Tweet"] = test\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_special\_characters)

p3=re.compile("https\*")

test\_tweets["Cleaned\_Tweet"] = [p3.sub ('', x) for x in test\_tweets["Cleaned\_Tweet"]]

p4 = re.compile (r"\d+")

train\_tweets["Cleaned\_Tweet"] = [p4.sub ('NUM', x) for x in train\_tweets["Cleaned\_Tweet"]]

p5=re.compile("@[A-Za-z0-9\_]+")

train\_tweets["Cleaned\_Tweet"] = [p5.sub ('', x) for x in train\_tweets["Cleaned\_Tweet"]]

train\_tweets["Cleaned\_Tweet"] = train\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_emojis)

train\_tweets["Cleaned\_Tweet"] = train\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_multiple\_spaces)

train\_tweets["Cleaned\_Tweet"] = train\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_urls)

train\_tweets["Cleaned\_Tweet"] = train\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_punctuations)

train\_tweets["Cleaned\_Tweet"] = train\_tweets["Cleaned\_Tweet"].apply(nfx.remove\_special\_characters)

p6=re.compile("https\*")

train\_tweets["Cleaned\_Tweet"] = [p6.sub ('', x) for x in train\_tweets["Cleaned\_Tweet"]]

#print(test\_tweets.head())

#print(train\_tweets.head())

#remove stop words

stop\_words=stopwords.words('english')

test\_tweets['Cleaned\_Tweet'] = test\_tweets['Cleaned\_Tweet'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop\_words)]))

train\_tweets['Cleaned\_Tweet'] = train\_tweets['Cleaned\_Tweet'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop\_words)]))

#print(test\_tweets.head())

#print(train\_tweets.head())

#Tokenization

vectorizer = CountVectorizer()

cleaned\_\_test\_tweet= test\_tweets["Cleaned\_Tweet"]

vectorized\_test\_tweets = vectorizer.fit\_transform(cleaned\_\_test\_tweet)

test\_tweets\_array = vectorized\_test\_tweets.toarray()

#print(vectorizer.get\_feature\_names\_out())

#print(vectorized\_test\_tweets.toarray())

#print(vectorized\_test\_tweets.shape)

cleaned\_test\_tweets=test\_tweets["Cleaned\_Tweet"]

cleaned\_train\_tweet= train\_tweets["Cleaned\_Tweet"]

vectorized\_train\_tweets = vectorizer.fit\_transform(cleaned\_train\_tweet)

#print(vectorizer.get\_feature\_names\_out())

#print(vectorized\_test\_tweets.toarray())

#print(vectorized\_test\_tweets.shape)

cleaned\_train\_tweets=train\_tweets["Cleaned\_Tweet"]

#TFIDF

tfidf=TfidfVectorizer()

tfidf\_train\_tweets=tfidf.fit\_transform(cleaned\_train\_tweet)

print(type(tfidf\_train\_tweets))

print(len(tfidf.vocabulary\_))

print(tfidf.vocabulary\_)

print(tfidf\_train\_tweets)

tfidf\_train\_tweets\_array = tfidf\_train\_tweets.toarray()

print(type(tfidf\_train\_tweets\_array))

print(tfidf\_train\_tweets\_array)

tfidf\_test\_tweets=tfidf.transform(cleaned\_\_test\_tweet)

print(type(tfidf\_test\_tweets))

print(len(tfidf.vocabulary\_))

print(tfidf.vocabulary\_)

print(tfidf\_test\_tweets)

tfidf\_test\_tweets\_array = tfidf\_test\_tweets.toarray()

print(type(tfidf\_test\_tweets\_array))

print(tfidf\_test\_tweets\_array.shape)

# Train Model

#assign training and testing

x\_test=tfidf\_test\_tweets\_array

x\_train=tfidf\_train\_tweets\_array

y\_train=train\_tweets['Label']

model=LogisticRegression(C=1,penalty='l1', solver='saga').fit(x\_train, y\_train)

print(model.predict\_proba(x\_test))

y\_predict = [int(p[1]>0.5)for p in model.predict\_proba(x\_test)]

print(y\_predict)

y\_predict\_train = [int(p[1]>0.5)for p in model.predict\_proba(x\_train)]

print(accuracy\_score(y\_train, y\_predict\_train))

print(len(model.coef\_[0]))

feature\_names = tfidf.get\_feature\_names()

for i,col in enumerate (feature\_names):

if np.abs(model.coef\_[0, i])>1e9:

print(feature\_names[i], ' - ', model.coef\_[0,i])

print(len(feature\_names))

#Classification Report

print(classification\_report(y\_train, y\_predict\_train))

print(confusion\_matrix(y\_train, y\_predict\_train))

1. David Sacks, “The China-Pakistan Economic Corridor-Hard Reality Greets BRI’s Signature Initiative,” Council on Foreign Relations, March 30, 2021, <https://www.cfr.org/blog/china-pakistan-economic-corridor-hard-reality-greets-bris-signature-initiative>. [↑](#footnote-ref-1)
2. Alex W. Palmer, “The Man Behind China’s Aggressive New Voice,” The New York Times Magazine, July 1, 2021, https://www.nytimes.com/2021/07/07/magazine/china-diplomacy-twitter-zhao-lijian.html [↑](#footnote-ref-2)
3. Palmer, “The Man Behind China’s Aggressive New Voice.” [↑](#footnote-ref-3)
4. Sacks, “The China-Pakistan Economic Corridor.” [↑](#footnote-ref-4)
5. Sacks, “The China-Pakistan Economic Corridor.” [↑](#footnote-ref-5)
6. Li Qianqian and Liu Yijun, “The China-Pakistan Economic Corridor: The Pakistani Media Attitudes Perspective,” Technology in Society, vol. 62 (August 2020): 2-5 <https://doi.org/10.1016/j.techsoc.2020.101303> [↑](#footnote-ref-6)
7. Sujan Chinoy, “CPEC: China’s designs, Pakistan’s Ambivalence and India’s Opposition, Hindustan Times, January 19, 2021, <https://www.hindustantimes.com/opinion/cpec-china-s-designs-pakistan-s-ambivalence-and-india-s-opposition-101611066633406.html> [↑](#footnote-ref-7)
8. DD News, “China Ignores India’s Sovereignty Concerns, Plans to extend Controversial CPEC to Afghanistan, DD News, May 24, 2021, <https://ddnews.gov.in/international/china-ignores-india%E2%80%99s-sovereignty-concerns-plans-extend-controversial-cpec-afghanistan> [↑](#footnote-ref-8)
9. Paras Ratna, “India’s CPEC Dilemma – To Participate or not to Participate,” East Asia Research Programme, <http://earp.in/en/indias-cpec-dilemma-to-be-or-not-to-be/> [↑](#footnote-ref-9)
10. Xiangming Chen, S.K. Joseph, and Hamna Tariq, "Betting Big on CPEC." *European Financial Review*. (February 10, 2018); 61-70. Accessed September 1, 2021, <http://www.europeanfinancialreview.com/?p=21658> [↑](#footnote-ref-10)
11. Christian Wagner, “The effects of the China-Pakistan Economic Corridor on India-Pakistan relations,”(SWP Comment, 25/2016). Berlin: Stiftung Wissenschaft und Politik -SWP- Deutsches Institut für Internationale Politik und Sicherheit. Accessed September 1, 2021, https://nbn-resolving.org/urn:nbn:de:0168-ssoar-46898-4 [↑](#footnote-ref-11)
12. David Sacks, “The China-Pakistan Economic Corridor-Hard Reality Greets BRI’s Signature Initiative,” *Council on Foreign Relation*s, March 30, 2021. Accessed September 1, 2021, <https://www.cfr.org/blog/china-pakistan-economic-corridor-hard-reality-greets-bris-signature-initiative>. [↑](#footnote-ref-12)
13. Alex W. Palmer, “The Man Behind China’s Aggressive New Voice,” The New York Times Magazine. (July 1, 2021). Accessed September 1, 2021, https://www.nytimes.com/2021/07/07/magazine/china-diplomacy-twitter-zhao-lijian.html [↑](#footnote-ref-13)
14. Palmer, “The Man Behind China’s Aggressive New Voice.” [↑](#footnote-ref-14)
15. Sacks, “The China-Pakistan Economic Corridor.” [↑](#footnote-ref-15)
16. Fakhar Hussain, Dr. MianSaeed Ahmad, Saadat Nawaz, Syed AqeelHaider, and Muhammad Atif, “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” *EEO.* 2021; 20(5): 3764-3770. Accessed September 1, 2021, Doi:10.17051/ilkonline.2021.05.414. [↑](#footnote-ref-16)
17. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-17)
18. Sacks, “The China-Pakistan Economic Corridor.” [↑](#footnote-ref-18)
19. Li Qianqian and Liu Yijun, “The China-Pakistan Economic Corridor: The Pakistani Media Attitudes Perspective,” Technology in Society, vol. 62 (August 2020): 2-5. Accessed September 1, 2021, <https://doi.org/10.1016/j.techsoc.2020.101303> [↑](#footnote-ref-19)
20. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-20)
21. Peter Cai, “Why India Distrusts China’s One Belt One Road Initiative,” *The Lowly Institute,* (September 2, 2016). Accessed September 1, 2021, https://www.lowyinstitute.org/the-interpreter/why-india-distrusts-chinas-one-belt-one-road-initiative [↑](#footnote-ref-21)
22. Sacks, “The China-Pakistan Economic Corridor.” [↑](#footnote-ref-22)
23. Li Qianqian “The China-Pakistan Economic Corridor: The Pakistani Media Attitudes Perspective,” 2-5. [↑](#footnote-ref-23)
24. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-24)
25. Sujan Chinoy, “CPEC: China’s designs, Pakistan’s Ambivalence and India’s Opposition” *Hindustan Times*, (January 19, 2021). Accessed September 1, 2021, <https://www.hindustantimes.com/opinion/cpec-china-s-designs-pakistan-s-ambivalence-and-india-s-opposition-101611066633406.html> [↑](#footnote-ref-25)
26. Paras Ratna, “India’s CPEC Dilemma – To Participate or not to Participate,” East Asia Research Programme. Accessed September 1, 2021, <http://earp.in/en/indias-cpec-dilemma-to-be-or-not-to-be/> [↑](#footnote-ref-26)
27. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-27)
28. Soumya, “China-led BRI and its Impact on India with Specific Reference to CPEC.” [↑](#footnote-ref-28)
29. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-29)
30. Wagner, “The effects of the China-Pakistan Economic Corridor on India-Pakistan relation.”  [↑](#footnote-ref-30)
31. Chinoy, “CPEC: China’s designs, Pakistan’s Ambivalence and India’s Opposition” [↑](#footnote-ref-31)
32. Arzan Tarapore, “The Crisis After the Crisis: How Ladakh will Shape India’s Competition with China,” Lowy Institute, (May 6, 2021). Accessed September 1, 2021, https://www.lowyinstitute.org/publications/crisis-after-crisis-how-ladakh-will-shape-india-s-competition-china [↑](#footnote-ref-32)
33. Soumya Awasthi, “China-led BRI and its Impact on India with Specific Reference to CPEC,” *Journal of International Affairs* 2, no. 1 (2018), accessed September 1, 202, <https://doi.org/10.3126/joia.v2i1.22592> [↑](#footnote-ref-33)
34. Ratna, “India’s CPEC Dilemma – To Participate or not to Participate.” [↑](#footnote-ref-34)
35. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-35)
36. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-36)
37. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-37)
38. Muhammad Ijaz Latif and Muhammad Tayyab Zia, “Strategic Dimensions of CPEC: Role of Regional and International Powers,” *Journal of Business and Social Review in Emerging Economies*, 6(4), 1561-1569 (2020). Accessed on September 1, 2-21, https://doi.org/10.26710/jbsee.v6i4.1507 [↑](#footnote-ref-38)
39. Latif and Zia, “Strategic Dimensions of CPEC,” 1561-1569 [↑](#footnote-ref-39)
40. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-40)
41. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-41)
42. Soumya, “China-led BRI and its Impact on India with Specific Reference to CPEC.” [↑](#footnote-ref-42)
43. Latif and Zia, “Strategic Dimensions of CPEC,” 1561-1569 [↑](#footnote-ref-43)
44. Latif and Zia, “Strategic Dimensions of CPEC,” 1561-1569 [↑](#footnote-ref-44)
45. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-45)
46. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-46)
47. Latif and Zia, “Strategic Dimensions of CPEC,” 1561-1569 [↑](#footnote-ref-47)
48. Soumya, “China-led BRI and its Impact on India with Specific Reference to CPEC.” [↑](#footnote-ref-48)
49. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-49)
50. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-50)
51. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-51)
52. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-52)
53. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-53)
54. Baluchistan is sometimes spelled Balochistan. [↑](#footnote-ref-54)
55. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-55)
56. Hafeez Ullah Khan. (2019). Regional Security Threat to CPEC: A Strategic Overview. Journal of the Research Society of Pakistan. VOL 56 Issue No 1 (January – June, 2019). https://www.researchgate.net/publication/335095537\_Regional\_Security\_Threats\_to\_CPEC\_A\_Strategic\_Overview [↑](#footnote-ref-56)
57. Hafeez Ullah Khan. (2019). Regional Security Threat to CPEC: A Strategic Overview. Journal of the Research Society of Pakistan. VOL 56 Issue No 1 (January – June, 2019). https://www.researchgate.net/publication/335095537\_Regional\_Security\_Threats\_to\_CPEC\_A\_Strategic\_Overview [↑](#footnote-ref-57)
58. Soumya, “China-led BRI and its Impact on India with Specific Reference to CPEC.” [↑](#footnote-ref-58)
59. Houreld, Katharine. (2015, April 20). China and Pakistan Launch Economic Corridor Plan Worth $45 Billion. Rueters. [China and Pakistan launch economic corridor plan worth $46 billion | Reuters](https://www.reuters.com/article/us-pakistan-china-idUSKBN0NA12T20150420) [↑](#footnote-ref-59)
60. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-60)
61. Wagner, “The effects of the China-Pakistan Economic Corridor on India-Pakistan relation.”  [↑](#footnote-ref-61)
62. Houreld, “China and Pakistan Launch Economic Corridor Plan Worth $45 Billion.” [↑](#footnote-ref-62)
63. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-63)
64. Wagner, “The effects of the China-Pakistan Economic Corridor on India-Pakistan relation.”  [↑](#footnote-ref-64)
65. Wagner, “The effects of the China-Pakistan Economic Corridor on India-Pakistan relation.”  [↑](#footnote-ref-65)
66. Soumya, “China-led BRI and its Impact on India with Specific Reference to CPEC.” [↑](#footnote-ref-66)
67. Wagner, “The effects of the China-Pakistan Economic Corridor on India-Pakistan relation.”  [↑](#footnote-ref-67)
68. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-68)
69. Cai, “Why India Distrusts China’s One Belt One Road Initiative” [↑](#footnote-ref-69)
70. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-70)
71. Soumya, “China-led BRI and its Impact on India with Specific Reference to CPEC.” [↑](#footnote-ref-71)
72. Wagner, “The effects of the China-Pakistan Economic Corridor on India-Pakistan relation.”  [↑](#footnote-ref-72)
73. Chen, et al. "Betting Big on CPEC," 61-70. [↑](#footnote-ref-73)
74. Soumya, “China-led BRI and its Impact on India with Specific Reference to CPEC.” [↑](#footnote-ref-74)
75. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-75)
76. Fakhar et al., “Challenges to China-Pak Economic Corridor (CPEC): The Indian Perspective,” 3764-3770. [↑](#footnote-ref-76)
77. Key words were determined based on themes identified in the Literature Review. Key words include CPEC, China-Pakistan, OBOR, BRI, BeltandRoad, Pakistan-China, Gwadar, and Kashgar. [↑](#footnote-ref-77)
78. A standard Twitter developer account was used for this research, precluding the advanced search criteria for countries as a whole. As a work around, the python script used geocode=”20.5937, 78.9629, 3000km” to match as closely as possible to India-origin tweets. [↑](#footnote-ref-78)
79. Atharva Mashalkar. Sentiment Analysis using Logistic Regression and Naïve Bayes. Towards Data Science. November 28, 2020. Accessed October 27, 2021, https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b [↑](#footnote-ref-79)
80. TFIDF was conducted using Python library Scikit-learn (Sklearn) CountVectorizer and TfidfVectorizer [↑](#footnote-ref-80)
81. Kuhn, Max, and Johnson, Kjell. Feature Engineering and Selection: A Practical Approach for Predictive Models. Milton: CRC Press LLC, 2019 (p.114-120). [↑](#footnote-ref-81)
82. In a study on movie reviews, supervised techniques performed with about 85% accuracy and unsupervised techniques performed at approximately 77% accuracy. Neri, Federico & Aliprandi, Carlo & Capeci, Federico & Cuadros, Montse & By, Tomas. (2012). Sentiment Analysis on Social Media. 10.1109/ASONAM.2012.164. [↑](#footnote-ref-82)
83. N. Mamgain, E. Mehta, A. Mittal and G. Bhatt, "Sentiment analysis of top colleges in India using Twitter data," 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016, pp. 525-530, doi: 10.1109/ICCTICT.2016.7514636. [↑](#footnote-ref-83)
84. Amina, Bibi & Azim, Tayyaba. (2019). SCANCPECLENS: A Framework for Automatic Lexicon Generation and Sentiment Analysis of Micro Blogging Data on China Pakistan Economic Corridor. IEEE Access. 7. 133876 - 133887. 10.1109/ACCESS.2019.2940528. [↑](#footnote-ref-84)
85. Anuja Nagpal. “L1 and L2 Regularization Methods.” Towards Data Science. (October 13, 2017). Accessed October 24, 2021, https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c [↑](#footnote-ref-85)
86. Santosh Kumar Pothabattula. “The Math and Intuition behind Logistic Regression.” Analytics Vidhya. August 25, 2019. Accessed October 24, 2021, https://medium.com/analytics-vidhya/a-complete-understanding-of-how-the-logistic-regression-can-perform-classification-a8e951d31c76 [↑](#footnote-ref-86)
87. Lasso regression adds the absolute value of magnitude of the coefficient as a penalty term to the loss function whereas Ridge regression adds the squared magnitude of the coefficient. Lasso shrinks the coefficient of the less important features, removing some features entirely, which serves well for cases with a high number of features. Anuja Nagpal. “L1 and L2 Regularization Methods.” Towards Data Science. (October 13, 2017). Accessed October 24, 2021, https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c [↑](#footnote-ref-87)
88. Santosh Kumar Pothabattula. “The Math and Intuition behind Logistic Regression.” Analytics Vidhya. August 25, 2019. Accessed October 24, 2021, https://medium.com/analytics-vidhya/a-complete-understanding-of-how-the-logistic-regression-can-perform-classification-a8e951d31c76 [↑](#footnote-ref-88)