Blockchain in Healthcare: a New Perspective from Social Media Data

Andrew Caietti

Follow this and additional works at: https://scholarworks.wm.edu/honorstheses

Part of the Other Applied Mathematics Commons

Recommended Citation
https://scholarworks.wm.edu/honorstheses/1728

This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
Blockchain in Healthcare: a New Perspective from Social Media Data

Andrew Caietti

Department of Mathematics
William & Mary
Williamsburg, VA 23187-8795, USA
Abstract

Blockchain as a technology has brought with it a wave of promises and expectations. After its successes in the financial sector, many potential new applications of the technology have been theorized across a variety of sectors. Blockchain’s application to healthcare stands out among these theories. Healthcare is a sector that views technological innovation under more scrutiny, so the introduction of blockchain into healthcare is a particularly unique implementation of the technology. Attempting to understand how blockchain is accepted in the healthcare industry is a difficult problem due to the nature of data associated with the sector. One avenue to understand how blockchain is viewed by this sector is through analysis of social media micro-blogging on the Twitter platform. By archiving a time series of tweets, important questions about how blockchain is viewed in healthcare can be addressed with the natural language processing technique of sentiment analysis. An ensemble of BERT models are identified as the best classifier with the given training data, and are further applied to a time series of tweets about blockchain in healthcare. This study analyzes healthcare perceptions of blockchain based on these results, and finds that the distribution of sentiment is largely positive. Examining the volume of tweets over time also indicates a massive increase in interest in the topic in 2018. Finally, when exploring how company accounts tweet compared to personal accounts, it is found that personal accounts produce slightly more positive tweets relative to company accounts. Thus, it is understood that healthcare perception of blockchain became consistently positive following 2017.
# Blockchain in Healthcare: a New Perspective from Social Media Data

*William & Mary Department of Mathematics*

Andrew Caietti

## Contents

1. **Introduction** .................................................. 5

2. **Sentiment Analysis** ............................................ 7
   2.1 Document vs. Sentence Level Analysis .......................... 7
   2.2 Related Work .................................................. 9
      2.2.1 Models .................................................. 9
      2.2.2 Binary vs. Multi-class Problem .......................... 10

3. **Data Acquisition** ............................................. 10
   3.1 Training and Testing Datasets .................................. 11
   3.2 Data Pre-Processing ........................................... 12
      3.2.1 Remove Duplicates ...................................... 13
      3.2.2 Hyperlink Remove ...................................... 13
      3.2.3 Remove User Mentions .................................. 13
      3.2.4 Lower-Case Text ........................................ 13
      3.2.5 Hashtags ................................................ 13
      3.2.6 Whitespaces ............................................. 14
      3.2.7 Miscellaneous Text Removal ............................. 14
   3.3 Feature Extraction ............................................ 14
      3.3.1 Term Frequency .......................................... 14
      3.3.2 Inverse Document Frequency ............................... 14
      3.3.3 TF-IDF .................................................. 15

4. **Classifier** .................................................. 15
   4.1 Models ........................................................ 16
      4.1.1 Naive Bayes ............................................ 16
      4.1.2 Support Vector Classifier ................................ 17
      4.1.3 Random Forest .......................................... 18
<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2 Model Evaluation</td>
</tr>
<tr>
<td>5 BERT</td>
</tr>
<tr>
<td>5.1 Background</td>
</tr>
<tr>
<td>5.2 Preparation</td>
</tr>
<tr>
<td>5.3 Binary Ensemble Model</td>
</tr>
<tr>
<td>5.4 Training Data for Binary Ensemble</td>
</tr>
<tr>
<td>5.4.1 BERT OVO vs. OVA Evaluation</td>
</tr>
<tr>
<td>5.5 Model Application</td>
</tr>
<tr>
<td>6 Results</td>
</tr>
<tr>
<td>6.1 General</td>
</tr>
<tr>
<td>6.1.1 Bots</td>
</tr>
<tr>
<td>6.2 Volume</td>
</tr>
<tr>
<td>6.3 Sentiment</td>
</tr>
<tr>
<td>6.4 Account Type Classification</td>
</tr>
<tr>
<td>7 Discussion</td>
</tr>
<tr>
<td>7.1 Post-hoc Work</td>
</tr>
<tr>
<td>7.2 Conclusion</td>
</tr>
</tbody>
</table>
List of Figures

1. Initial train and test split logic. .................................................. 19
2. Example of the K-Fold logic where k=4. .................................... 19
3. Binary Ensemble Logic ................................................................. 25
4. Percent of total accounts broken down into bots, not bots, and N/A. ... 29
5. Volumes of tweets over time ......................................................... 30
6. Percent distribution of classes in the data set. .......................... 32
7. The percent distribution of classes aggregated by month, with a constant line at 50% 33
8. Box plots for Positive class over Year .......................................... 33
9. Box plots for Positive class grouped before versus after 2018 ........ 33
10. Box plots for Neutral class over Year ......................................... 34
11. Box plots for Neutral class grouped before versus after 2018 ......... 34
12. Box plots for Negative class over Year ....................................... 34
13. Box plots for Negative class grouped before versus after 2018 ...... 34
14. Percent distribution of classes broken down over time aggregated by month. ....................... 35
15. Percent of company versus personal accounts. .......................... 37
16. Total tweets of company versus personal accounts. .................... 37
17. Sentiment class distribution of total tweets aggregated by account type. ......................... 38
18. Sentiment class distribution of total tweets aggregated by account type and excluding the neutral class. 39

List of Tables

1. Table of literature and the methodologies followed ....................... 9
2. Manually relabeled tweets examples .......................................... 11
3. Examples of tweets in the training dataset associated with each class .................... 12
4. Example probability values from an OVA ensemble .................... 26
5. Example probability values from an OVO ensemble .................... 26
6. Lines of best-fit by sentiment class. ......................................... 35
1 Introduction

Blockchain has seen a dramatic rise to relevance in the past decade. Blockchain applications in healthcare are becoming more recognized as the technology is realized in other sectors. Blockchain’s proven ability to provide appealing innovation such as a distributed ledger, immutability, and decentralized storage appeals to innovators in the healthcare sector (Mackey et al., 2019). Yet, with the innovation comes challenges in application and skepticism of its applicability. One such healthcare-specific challenge to adopting new technologies is the Health Insurance Portability and Accountability Act of 1996 (HIPAA). HIPAA requires factors such as interoperability, record sharing, and authentication as legal requisites for emerging technology (McGhin et al., 2019). Legal requirements considered, the healthcare sector characteristically is also resistant to changing current practices (Sligo et al., 2017). Given notable hesitations, literature sees opportunity in blockchain for healthcare.

Blockchain technology has the ability to improve patient safety through device tracking and pharmaceutical traceability (Bell et al., 2018). The concept of tracing is similar to tracking patient data. HIPAA sets strict requirements for how patient data is handled by all participants in healthcare (Assistance, 2003), making interoperability, the exchange of information between key players, in the healthcare sector challenging yet important (Iroju et al., 2013). While blockchain can provide private and auditable data sharing (Theodouli et al., 2018), concerns surrounding scalability and security are important considerations for blockchain’s future in healthcare (Abu-Elezz et al., 2020). A notable hurdle to blockchain’s adaptation is the “social acceptance of blockchain technology” (Abu-Elezz et al., 2020; Cao et al., 2019; Kaur et al., 2018; Leeming et al., 2019; Patel, 2019; McGhin et al., 2019) particularly in its applications in healthcare as an important piece to understanding the puzzle of blockchain’s future in the healthcare sector. Many authors have pointed out this direction as a research gap.

Understanding the social acceptance of blockchain technology is particularly important, thanks to how hyped blockchain technology is. The hype stems from the many promises of innovation that blockchain technology has made. Thus, many companies have been created to bring the business value of blockchain technology into various sectors. For a company built around a new technology, said company wants to see that technology succeed and be viewed in a positive light. Positive views of the technology can generate positive business for these companies. This simple line of logic can be drawn towards blockchain and these newer organizations built around generating revenue by bringing the business value of blockchain to various sectors. Since these companies have an identifiable vested interest in blockchain technology succeeding, their sentiments may differ from the sentiments of accounts that are not representative of a company, or non-company accounts. Understanding how company sentiment may differ from these non-company accounts is particularly important to understanding the true public perception of blockchain in healthcare. While companies have a vested interest, non-company Twitter accounts may not necessarily have the
same investiture into the technology. These individuals when talking about blockchain in healthcare could speak to a variety of subtopics, from commenting on a company’s activity with the technology to expressing general opinion toward blockchain itself and its relation in healthcare. Thus, while a company may have vested interest in the success of the technology, a non-company participant in the conversation may express a different sentiment. Given the potential for these two participants in the public conversation of blockchain in healthcare to have differing sentiments, it is important then to identify these groups and understand how their sentiments differ.

To better understand this problem, I can use Twitter data. Twitter is a popular trademarked social media platform where users from across the globe are able to produce textual messages of up to 280 characters about any host of topics. Having a daily active user base of 192 million users\(^1\) makes Twitter a popular medium to explore applications of sentiment analysis (Mittal and Goel, 2012; Wu et al., 2015; Rouhani and Abedin, 2019; Daniel et al., 2017; Kraaijeveld and Smedt, 2020), as the volume of textual data available can span a broad range of topics and fields. One such field is the emerging technology of Blockchain. Emerging in 2008\(^2\), blockchain saw its prominence grow with its use in the cryptocurrency Bitcoin. Yet the applications of blockchain as a new technology have been explored in various sectors beyond the financial sector (Miraz and Ali, 2018; Tasatanattakool and Techapanupreeda, 2018), one such sector being healthcare (Siyal et al., 2019).

Given the notable uncertainty and potential in blockchain’s application to the healthcare sector, it is important to understand people’s sentiment toward this technology, as this can improve the general understanding of how people perceive the applicability of blockchain in healthcare (Kuo et al., 2017). In this paper, I identify the most accurate classifier for this specific task, a One-Versus-One (OVO) approach of Bidirectional Encoder Representations from Transformers (BERT) models, then apply this model to Twitter data to classify textual sentiment to address these questions.

- What is the social acceptance of blockchain in healthcare?
- How does a company having a vested interest affect the sentiment of blockchain in healthcare?
- Does the tweet sentiment differ between company and non-company account tweets?

From its applications, sentiment is seen as an indicator of decision making (Lerner et al., 2015). Given the derived value in understanding public sentiment surrounding various topics, it is important to understand public opinion of blockchain in healthcare (Abu-Elezz et al., 2020) as an emerging technology in the healthcare space through Twitter data. Understanding this sentiment can provide insight into how the technology is being received in

\(^2\)https://www.economist.com/briefing/2015/10/31/the-great-chain-of-being-sure-about-things
the healthcare sector, and this can be accomplished through sentiment analysis.

2 Sentiment Analysis

Sentiment Analysis is a form of Natural Language Processing, defined specifically to be the task of inferring the opinion expressed in a given document. The task has become dramatically more popular since its inception, as the ability to identify opinions in textual data can yield important information in a variety of applications. Understanding sentiment can drive decision making. The financial sector has realized this and the importance of understanding public opinion by incorporating this information with market data to predict future stock prices (Batra and Daudpota, 2018; Mittal and Goel, 2012; Khedr et al., 2017). Companies needing to understand public opinion and perception of a given product or brand often turn to sentiment analysis across various mediums, with social media being a popular application (Ghiassi et al., 2013; Vidya et al., 2015; Chamlertwat et al., 2012).

The sentiment of a textual document can be defined in a multitude of ways, yet is generally seen as a set of labels indicating a range of emotions. Most commonly, this is seen as a binary problem of classifying documents as either positive or negative. Yet given the broad definition of the task, the classes a given classifier discerns are flexibly determined by the finer definition of the application of sentiment analysis a user identifies. Beyond the binary positive and negative class sentiment analysis tasks, one may encounter multi-class sentiment analysis. These applications are considered fine-grain variations of sentiment analysis, and can range in the number of classes from a hedonometric scale of sentiment intensity to the addition of classes to the binary task, neutral being a prime example. Yet as the number of classes in a given document increases beyond the binary case, the complexity of the problem increases and the ability of conventional classification techniques to perform in these multi-class scenarios becomes more challenging.

2.1 Document vs. Sentence Level Analysis

In framing the problem of sentiment analysis, there are various factors that affect the problem that must be considered. There are different levels of granularity in which the problem of sentiment analysis can be addressed. The first methodology is Document-level sentiment classification. This entails the processing of a given textual document in its entirety, then the labeling of said document with a sentiment classification. So, given a set of text documents, \( D \), each individual document \( d \in D \) is processed and a classification, \( c_i \) is assigned (Liu et al., 2010). Any given user authoring a document \( d \in D \) is then assumed to be presenting their personal views, or opinion (Ravi and Ravi, 2015). This underlying assumption that the opinion reflected by a document \( d \) carries implication toward the opinion of the author of the document is important to understanding the public perception of a given topic at a macro scale.
So, an example may be considered: "Blockchain is proven to be successful in application and drives value for companies. Healthcare adoption must be scrutinized." In this instance, I see strong positive terminology associated with blockchain, identifying the technology as successful and a driver of value. So, despite the second sentence indicating negative sentiment, the general sentiment of the document is positive since there is more polarity in positive speech than neutral or negative speech. The important distinction here is that classification occurs at the document level and, despite the manual determination between sentences, I only classify the document overall and not individual sentences.

Another methodology is Sentence-level sentiment classification. This is the process of classifying the sentiments of sentences contained within a document. So, for each document \( d_i \in D \) for \( i = 1, \ldots, n \), sentences \( s_j \in d_i \) for \( j = 1, \ldots m \) are each individually evaluated and further classified into a given sentiment class (Liu et al., 2010). This methodology follows a more granular approach, and can then extrapolate toward classifying the document by understanding the polarity of sentiment for each sentence.

Another example may be considered: "Blockchain is a hopeful technology in other sectors, and has potential in finance. However, Blockchain in healthcare is impossible and has no future." In this case, the document \( d_i \) is split into two sentences \( s_1, s_2 \in d_i \), where:

\[
\begin{align*}
s_1 & := \text{Blockchain is a hopeful technology in other sectors, and has potential in finance.} \\
\text{(1)} \\
s_2 & := \text{However, Blockchain in healthcare is impossible and has no future.} \\
\text{(2)}
\end{align*}
\]

The first sentence is classified as positive, while the second sentence is classified as negative. The polarity of each sentence’s classification could then be considered in determining if the document \( d_i \) is positive or negative. In this case, the polarity \( s_2 \) is notably more negative than the positive polarity of \( s_1 \). So, the given document is classified as negative.

For purposes of this study, document-level sentiment analysis is conducted. This is chosen in the context of the processed data being Twitter data. Each tweet can consist of a maximum of 280 characters. Furthermore, Twitter is a social media platform, so each tweet within the given length constraints are less structured than, say, sentences in an academic article. Thus, given the relative unstructured nature of tweets and the constraints on length of each tweet, sentiment is determined on a document level where a given \( d_i \in D \) is defined to be an individual tweet.
2.2 Related Work

2.2.1 Models

Sentiment analysis to understand blockchain in healthcare is not frequent in the literature. Thus, to examine how sentiment analysis has been performed, a broader viewpoint can be taken. This allows us to understand how sentiment analysis generally has been conducted as well as in the more domain specific areas such as sentiment analysis on cryptocurrency in finance, and methodologies followed in the healthcare sector. Viewing the literature within this scope, I see a variety of methodologies followed. This is reflected in Table 1 below.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Approach</th>
<th>Classifier</th>
<th>Classes</th>
<th>Data</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Abraham et al., 2018)</td>
<td>lexicon</td>
<td>VADER</td>
<td>positive or neutral or negative</td>
<td>tweet data</td>
<td>finance</td>
</tr>
<tr>
<td>(Colianni et al., 2015)</td>
<td>machine learning</td>
<td>LSVM, Logistic Regression, Naive Bayes</td>
<td>positive or neutral or negative</td>
<td>tweet data</td>
<td>finance</td>
</tr>
<tr>
<td>(Albrecht et al., 2019)</td>
<td>lexicon</td>
<td>SentiStrength</td>
<td>quantitative ranking</td>
<td>misc dataset + tweet data</td>
<td>finance</td>
</tr>
<tr>
<td>(Shresthabai et al., 2020)</td>
<td>machine learning</td>
<td>LSTM (NN)</td>
<td>positive or negative</td>
<td>misc dataset</td>
<td>finance</td>
</tr>
<tr>
<td>(Agarwal, 2020)</td>
<td>lexicon</td>
<td>VADER</td>
<td>positive or neutral or negative</td>
<td>financial news headlines</td>
<td>finance</td>
</tr>
<tr>
<td>(Dobele, 2019)</td>
<td>machine learning</td>
<td>LSTM, SVM</td>
<td>positive or neutral or negative</td>
<td>news articles</td>
<td>finance</td>
</tr>
<tr>
<td>(Bonnes et al., 2018)</td>
<td>lexicon + machine learning</td>
<td>Linear Regression</td>
<td>positive or negative</td>
<td>articles</td>
<td>finance</td>
</tr>
<tr>
<td>(Li et al., 2019a)</td>
<td>machine learning</td>
<td>Regression Techniques, SVM</td>
<td>positive or neutral or negative</td>
<td>tweet data</td>
<td>finance</td>
</tr>
<tr>
<td>(Clark et al., 2018)</td>
<td>machine learning</td>
<td>Logistic Regression, Convolutional Neural Network</td>
<td>positive or negative</td>
<td>tweet data</td>
<td>healthcare</td>
</tr>
<tr>
<td>(Wu et al., 2015)</td>
<td>machine learning</td>
<td>SVM</td>
<td>positive or negative</td>
<td>tweet data</td>
<td>healthcare</td>
</tr>
<tr>
<td>(Islam and Sultana, 2018)</td>
<td>machine learning</td>
<td>SVM, Naive Bayes, Random Forests</td>
<td>positive or negative</td>
<td>misc dataset</td>
<td>general sentiment analysis</td>
</tr>
<tr>
<td>(Asghar et al., 2013)</td>
<td>lexicon + machine learning</td>
<td>SentiWordNet, Naive Bayes</td>
<td>positive or negative</td>
<td>misc dataset</td>
<td>healthcare</td>
</tr>
<tr>
<td>(Asghar et al., 2014)</td>
<td>lexicon</td>
<td>SentiWordNet</td>
<td>positive or neutral or negative</td>
<td>misc dataset</td>
<td>healthcare</td>
</tr>
<tr>
<td>(Greaves et al., 2013)</td>
<td>machine learning</td>
<td>Naive Bayes, Decision Trees, SVM</td>
<td>quantitative ranking</td>
<td>survey data</td>
<td>healthcare</td>
</tr>
<tr>
<td>(Bacco et al., 2020)</td>
<td>machine learning</td>
<td>SVM, BERT</td>
<td>positive or negative</td>
<td>misc dataset</td>
<td>healthcare</td>
</tr>
<tr>
<td>(Kamwombe et al., 2019)</td>
<td>lexicon</td>
<td>textblob, VADER</td>
<td>positive or neutral or negative</td>
<td>misc dataset</td>
<td>healthcare</td>
</tr>
<tr>
<td>(Schneider et al., 2020)</td>
<td>lexicon + machine learning</td>
<td>Naive Bayes, SVM, Decision Tree</td>
<td>positive or negative</td>
<td>misc dataset</td>
<td>healthcare</td>
</tr>
</tbody>
</table>

Table 1: Table of literature and the methodologies followed

Upon review of the literature, both machine learning and lexical methods are performed. An analysis of Breast Cancer treatment experiences (Clark et al., 2018) utilized a combined approach of a logistic regression with a convolutional neural network to classify Twitter data on a hedonometric scale. Wu et al. (2015) utilized support vector machines to classify adverse drug reactions from tweet data. The use of support vector machines is quite frequent in literature (Wu et al., 2015; Islam and Sultana, 2018; Dobele, 2019; Bacco et al., 2020). Broadly, Table 1 shows how models vary from Naive Bayes to Support Vector Machines (SVM) to Long Short Term Memory neural networks, with SVM implementations being most frequent. Yet few pieces in the literature have utilized an ensemble of classifiers, namely they lack applications of BERT, in a One-Versus-One approach to classify sentiment in blockchain and healthcare. Thus, this paper shall explore the intersection of blockchain in healthcare and provide a unique approach as a study of applicability of this sentiment analysis technique.
2.2.2 Binary vs. Multi-class Problem

In terms of sentiment analysis, the problem can be viewed from a binary or multi-class perspective. From a binary, two class perspective, sentiment analysis is treated as a "class" or "not class" task (Clark et al., 2018; Islam and Sultana, 2018; Sureshbbhai et al., 2020). From a multi-class perspective, applications of sentiment analysis can range from a three class application (Li et al., 2019a; Dobele, 2019; Agarwal, 2020), to hedonometric scales as seen in (Greaves et al., 2013). For the purposes of our study, I will view the sentiment analysis task in terms of three classes positive, neutral, and negative. This focus is taken in order to get a general understanding of how a public forum such as Twitter views a technology entering the healthcare field, where understanding instances of positivity and negativity is meaningful enough to answer the questions I pose. Beyond this, a significant number of tweets can be void of polarity. For example, a company tweeting about a partnership with another company in blockchain and healthcare is neither positive nor negative. In this case, the tweet must be identified as neutral. So, using neutrality as a void of positive or negative sentiment is particularly important for the purposes of this project.

From the literature, it is further found that sentiment analysis to understand the healthcare sector acceptance of blockchain has not been extensively explored. This is particularly important, since blockchain as a technology has received a lot of “hype” since its conception and application in the cryptocurrency space (Carson et al., 2018), and more recently in its application in the healthcare space (El-Gazzar and Stendal, 2020). Thus, given the outlined need for a better understanding of social acceptance of blockchain technology, as outlined in the introduction, and the current lack of research in the area, this study shall further explore how blockchain in healthcare is perceived.

3 Data Acquisition

For this study, our corpus is composed of documents, tweets, which are up to 280 character text documents from the social media site Twitter. To acquire these tweets, this study utilized a Twitter developer account to access the Twitter API through a Python environment using the tweepy library (Roesslein, 2020). The Twitter API enables historical access to tweets by passing the ID of each tweet, then receiving a tweet object response from the API. To access Twitter IDs, a scraper was utilized in a Python environment. This scraper allowed for a Python script to be written to gather tweet IDs from January 2016 to January 2021, based on the keyword search of "blockchain healthcare." An example of a Twitter ID can be seen in the responses from the scraper implementation in the form of a URL, where USERNAME is a given Twitter username and the tweet ID observed is "692121578050600966".

https://Twitter.com/USERNAME/status/692121578050600966 (3)

Each URL is processed and the IDs are stored in a comma separated values (CSV) text file. These tweet IDs are finally passed to the Twitter API, which returns a tweet object, storing metadata associated with each tweet. These
objects are processed in JavaScript Object Notation, or JSON, format, and the necessary fields are extracted and stored. Based on this keyword search above, the resulting data set is composed of 207,287 tweets from 45,579 unique accounts.

3.1 Training and Testing Datasets

The training and test data set of 7570 tweets is taken as a subset of the main data set. For purposes of this study, this training data set is then labeled by using the Valence Aware Dictionary and Sentiment Reasoner (VADER) (Hutto and Gilbert, 2014). Following this, the training data set is left with 2586 positive, 2586 neutral, and 2398 negative tweets. These values account for the possibility of any duplicate tweets left within the data set, as duplicates can introduce bias into the model evaluation parameters referenced later.

<table>
<thead>
<tr>
<th>Text</th>
<th>VADER Label</th>
<th>Re-Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 3 #blockchain-based healthcare companies to watch in 2017</td>
<td>positive</td>
<td>neutral</td>
</tr>
<tr>
<td>Digital Health Pass powered by IBM’s Blockchain can transform healthcare and life sciences globally</td>
<td>neutral</td>
<td>positive</td>
</tr>
<tr>
<td>Real world blockchain applications in healthcare you probably missed via URL</td>
<td>negative</td>
<td>neutral</td>
</tr>
</tbody>
</table>

Table 2: Manually relabeled tweets examples

Following the automated classification using VADER, the data set is then re-examined manually, where class labels are manually confirmed or overridden. Uniformity is important in this process. The tweets were methodically examined by identifying misclassifications while also re-classifying a tweet if it is deemed incorrect in the context of blockchain and healthcare. Examples of a label being change are listed in Table 2. In the third row, for example, the tweet is labeled first by VADER as negative. Yet, upon reading, the tweet portrays a neutral sentiment as it is more so sharing a news article than it is expressing an opinion about any given topic. Thus, this label is changed from negative to neutral. A sample of tweets belonging to each class included in the training data set are listed in the table below.
Here’s a look at the new technology and the idea of this company, I loved everything in this crypto world, you will learn and understand the benefits!

Does #Blockchain have a place in #healthcare? (Definitely think so...at least for #publichealth). @forbes

Transforming the #healthcare systems and operations through #blockchain is something to look forward to because blockchain is the answer to solving all the irregularities in the system!

@USER Smart Contracts and Blockchain powered health information network (HIN) can fundamentally re-engineer the current healthcare systems and networks.

Follow the link and read our article How Blockchain Technology Will Transform Healthcare in 2018

Beyond Obamacare: Blockchain and the Future of Healthcare - #fintech #bitcoin #finance by #kingofpayments

Blockchain – How can it impact the healthcare industry?

Hyperledger sets up #blockchain working group for healthcare industry - EconoTimes #fintech

The healthcare blockchain claim "patients can control access to their data", i.e. restrict anyone incl provider from access to it: this one’s been touted since late 2016.

Are there any working implementations? Not "well you could try approach x", but existing present-day code?

Here’s a not-so-positive outlook on #healthcare trends in 2019. In summary, #AI, #blockchain, all big tech co & disruptors are collectively going to disappoint everyone.

A train-wreck of an article trying to make the case for blockchain in healthcare.

With every new technology, there is an initial bubble phase: 99% of the dot-com stocks disappeared, but a handful went to $1 trillion market caps. Mostly scam artist in the blockchain

<table>
<thead>
<tr>
<th>Text</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>#medchain #healthcare #blockchain here’s a look at the new technology and the idea of this company.</td>
<td>positive</td>
</tr>
<tr>
<td>I loved everything in this crypto world, you will learn and understand the benefits!</td>
<td>positive</td>
</tr>
<tr>
<td>Does #Blockchain have a place in #healthcare? (Definitely think so...at least for #publichealth). @forbes</td>
<td>positive</td>
</tr>
<tr>
<td>Transforming the #healthcare systems and operations through #blockchain is something to look forward to because blockchain is the answer to solving all the irregularities in the system!</td>
<td>positive</td>
</tr>
<tr>
<td>@USER Smart Contracts and Blockchain powered health information network (HIN) can fundamentally re-engineer the current healthcare systems and networks.</td>
<td>positive</td>
</tr>
<tr>
<td>Follow the link and read our article How Blockchain Technology Will Transform Healthcare in 2018</td>
<td>neutral</td>
</tr>
<tr>
<td>Beyond Obamacare: Blockchain and the Future of Healthcare - #fintech #bitcoin #finance by #kingofpayments</td>
<td>neutral</td>
</tr>
<tr>
<td>Blockchain – How can it impact the healthcare industry?</td>
<td>neutral</td>
</tr>
<tr>
<td>Hyperledger sets up #blockchain working group for healthcare industry - EconoTimes #fintech</td>
<td>neutral</td>
</tr>
<tr>
<td>The healthcare blockchain claim &quot;patients can control access to their data&quot;, i.e. restrict anyone incl provider from access to it: this one’s been touted since late 2016.</td>
<td>negative</td>
</tr>
<tr>
<td>Are there any working implementations? Not &quot;well you could try approach x&quot;, but existing present-day code?</td>
<td>negative</td>
</tr>
<tr>
<td>Here’s a not-so-positive outlook on #healthcare trends in 2019. In summary, #AI, #blockchain, all big tech co &amp; disruptors are collectively going to disappoint everyone.</td>
<td>negative</td>
</tr>
<tr>
<td>A train-wreck of an article trying to make the case for blockchain in healthcare.</td>
<td>negative</td>
</tr>
<tr>
<td>With every new technology, there is an initial bubble phase: 99% of the dot-com stocks disappeared, but a handful went to $1 trillion market caps. Mostly scam artist in the blockchain</td>
<td>negative</td>
</tr>
</tbody>
</table>

**Table 3:** Examples of tweets in the training dataset associated with each class

### 3.2 Data Pre-Processing

An important step of natural language processing (NLP) is the pre-processing of raw text data. The pre-processing of raw text data is the process in which text is cleaned and prepared for classification by a classifier model. This is particularly important because text from online sources, especially social media sites, is noisy and can be unintuitive to the non-human eye. Processing text data also assists in reducing the difficulty of the problem, since each word in a document represents a dimension. By reducing the noise and dimensions of the problem, this simplifies the classification process and consequentially improves the classifier accuracy. Thus, various steps in the pre-processing phase are followed as seen below.
3.2.1 Remove Duplicates

In some instances, user accounts will tweet the same tweet multiple times. For the training data, every duplicate instance of a tweet is removed. This is done not only to improve the efficiency of the model in training, but also to avoid a bias in the testing set. If multiple duplicated tweets are included in the test set, then the classifier will correctly identify the classes of those tweets and inflate the evaluation metrics and vice versa.

3.2.2 Hyperlink Remove

Hyperlinks are URLs that can be included in the raw text of each tweet. Each tweet obtained from the Twitter API contains at least one hyperlink, linking to itself, and is thus removed. Any further hyperlinks included in the tweet, with common examples being of images, links to news articles, or websites, are also removed from the tweet text.

3.2.3 Remove User Mentions

User mentions are commonly seen in tweets as a method to alert other users to a tweet. This is observed in the form of the "at-symbol" (@) followed by a string of characters representing a Twitter username, an example being "@POTUS". User mentions carry little by way of value for sentiment analysis, and are thus fully removed from the raw tweet text.

3.2.4 Lower-Case Text

Text from social media can show sentiment through a variety of means beyond the words chosen. For the purpose of this study, information conveyed through the meaning of each word is most important. For example, "Positive GROWTH" should, for purposes of this study, carry the same sentiment value as "positive growth", and as such must be interpreted as being the same words. Thus, each character of each word is converted to the lower case, so as to avoid the misconstruing of meaning of words.

3.2.5 Hashtags

Hashtags can contain sentiment value, and thus should not be rejected outright. For example, "#failure" can contain strong negative implications in the context of a tweet, and thus should not be excluded from said tweet. However, the pound sign, #, is removed from the tweet so as to reduce the dimensionality of the classification problem, while keeping the sentiment value of the hashtag.
3.2.6 Whitespaces
On occasion, tweets encountered in the data set had excessive whitespaces in the tweet itself. These whitespaces carry no sentiment value and thus are removed from relevant tweets.

3.2.7 Miscellaneous Text Removal
Various other instances of miscellaneous symbols had been found in the tweet text, and were removed from each tweet. Examples of these include expressions such as the new-line, \n, and other characters such as â or €.

3.3 Feature Extraction
Feature Extraction is the step in sentiment analysis, in which one examines textual information and identifies important features for consideration by the classifier. There are various forms of feature extraction in NLP; however, Term Frequency Inverse Document Frequency (TFIDF), a Bag of Words (BOW) approach, has been previously identified as a popular methodology in terms of efficiently retrieving information from text (Eklund, 2018). It is utilized in the case of this study due to its ease of implementation relative to the performance provided when compared to other feature extraction methods, namely word embedding (Eklund, 2018; Lilleberg et al., 2015).

TF-IDF is a statistic measuring the importance of a word for a document relative to the occurrence of that word in the corpus. In the case of applying this statistic to tweets, TF-IDF creates sparse vectors that weight the words in a tweet based on their frequency within the tweet itself, relative to their frequency of occurrence in the corpus of tweets. The TF-IDF value is calculated for each unique word in the documents evaluated, and represented by a sparse matrix of values corresponding to each word. This value is calculated in parts.

3.3.1 Term Frequency
Term Frequency (TF) is defined to be the frequency of occurrence of a word in a given document. In the context of this study, TF is the frequency of occurrence of a word in a given tweet. This value is calculated for each word in each document, \( w_i \in d \) where \( d := \{ w_i : i = 1, 2, \ldots, m \} \) and \( m \) is the total words in a document, by the formula:

\[
tf(w_1, d) = \frac{N(w_1)}{\sum_{i=1}^{m} N(w_i)} \quad (4)
\]

\[
N(w_i) = \text{count of a word in a document} \quad (5)
\]

\[
w_1 = \text{a unique word in a document} \quad (6)
\]

3.3.2 Inverse Document Frequency
Inverse Document Frequency (IDF) is defined to be the relative frequency of occurrence of a word in the given corpus of documents. In the context of this study, IDF is considered to be the relative frequency of occurrence
of a tweet in the given data set of tweets. This value is calculated for the corpus by defining each document, $d_j \in D := \{d_j : j := 1, 2, \ldots n\}$ where $n$ is the number of documents in the corpus, by the formula:

$$idf(w_1, D) = \log\left(\frac{N(d \in D)}{\sum_{j=1}^{n} N(w_1 \in d_j)}\right)$$

(7)

$$N(d \in D) = \text{count of total tweets in corpus}$$

(8)

$$N(w_1 \in d_j) = \text{count of total documents with word } w_1 \text{ in it}$$

(9)

3.3.3 TF-IDF

Together, the TF-IDF statistic is calculated by the following formula.

$$tf-idf(w, d, D) = tf(w, d) \cdot idf(w, D)$$

(10)

The statistic will effectively capture the weight of given words in their association with a class in a data set. For instance, if the word “improve” is commonly associated with the positive class, then TF-IDF is able to represent the weight of “improve” as associated with said class. So, if a new document is evaluated by the TF-IDF algorithm and “improve” is encountered, this will influence the probability of a positive class label being assigned to the document.

To utilize this statistic, this study leveraged the Python library SKLearn for calculation (Pedregosa et al., 2011).

4 Classifier

When considering the problem of deriving sentiment from textual data, classification models, known as classifiers, are particularly suited to the task. These classifiers are statistical models that analyze textual information and infer labels known as classes, based on the data.

Classification problems vary, and there is not a one-model-fits-all solution to be applied to a given text classification problem. Popular data sets utilized for training statistical models in text classification tasks in machine learning are generally based around topics that can have polarizing opinions. Some examples of these include but are not limited to product reviews, customer experiences, and politics (Feldman, 2013). The clear polarization can be seen in the documents included in the respective corpus for these topics, and allows a model to more clearly identify when a class label applies to a text document.

The focus of this study is on the problem of labeling tweet data that does not have the same strength in polarity between documents. Since blockchain is an emerging technology with potential to innovate in the healthcare sector, Twitter data does not reflect the same polarization that a movie review may possess. So, the problem of classifying tweet sentiment shifts. There is a combination of companies and personal users tweeting into this space. Tweets can range from company announcements about an event in the blockchain space, to headlines about company activity in adopting blockchain in healthcare, to individual reactions to the potential future or shortcomings of blockchain.
Understanding if blockchain is viewed by social media as a positive or negative addition into the healthcare sector yields tweets that exhibit a more news-oriented tone. The data set this study has mined and produced must then be considered under various different models to understand what classifier is most applicable to this text classification task.

4.1 Models

For purposes of identifying a diverse baseline of classifiers to understand what model works best for our task, various machine learning models are implemented using the SKLearn Python library (Pedregosa et al., 2011). Each model is implemented “out-of-the-box,” meaning the baseline performance of each model is observed without any hyper parameter tuning where applicable. The caveat to this is when Support Vector Machines are considered. Support Vector Machines are identified as frequently used for text classification tasks from the literature, as seen in Table 1. Thus, the implementation of this model underwent hyper parameter tuning. This was chosen to have a diverse baseline of models to best understand how different models can perform in this classification problem, while also understanding how one of the most common methods utilized in literature performs against these baselines when fully implemented.

4.1.1 Naive Bayes

Naive Bayes classifiers are built upon Bayes Theorem, entailing that the classifier assumes independence between each predictor. So, given the occurrence of a given predictor, the probability of the document being labeled a given class is calculated. This is done through the application of Bayes Theorem, where \( c \) is a given class, and \( x_1, x_2, \ldots, x_n \) is a given set of predictors:

\[
P(c|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|c)P(c)}{P(x_1, x_2, \ldots, x_n)}
\]  

(11)

Definitions of the probability of the likelihood of class, \( P(x_i|c) \), can give rise to various implementations of the Naive Bayes algorithm. In the implementation of a Gaussian Naive Bayes classifier, this entails that the likelihood probability is calculated with a maximum likelihood estimated variance, \( \sigma_y \), and mean, \( \mu_y \), as seen below:

\[
P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma^2_y}} e^{-\frac{(x_i - \mu_y)^2}{2\sigma^2_y}}
\]  

(12)
The Bernoulli Naive Bayes classifier assumes the data used for classification follows a Bernoulli distribution. The class likelihood probability be calculated by:

\[ P(x_i|c) = P(i|c)x_i + (1 - P(i|c))(1 - x_i) \]  

(13)

A Multinomial Naive Bayes classifier is also explored. This method assumes a multinomial distribution of input data. This method works particularly well when the input data is discrete, so using the TF-IDF statistic outlined above as our input data is particularly suited for this classifier. The conditional probability \( P(x_i|c) \) is calculated as:

\[ P(x_i|c) = \frac{\text{count}(x_i, c) + \alpha}{\text{count}(c) + \alpha \cdot n} \]  

where

\[ \alpha = \text{smoothing parameter (set to 1)}, \quad \text{and} \]  

\[ n = \text{number of words in vocabulary}. \]  

(14)

(15)

(16)

For purposes of implementation, each of the above variants of the Naive Bayes algorithms are implemented and evaluated.

4.1.2 Support Vector Classifier

Support Vector Machines address the classification task by constructing a hyper plane to divide the dimensional space of the input data. Classifications are then identified based on a datum’s position in the space relative to the hyper plane. Since this problem is a multi-class problem, the classifier can be configured in a “One-Versus-Rest” (OVA) or “One-Versus-One” (OVO) method. These two approaches pertain to how the support vector machines are configured for application to a multi-class task, as a single support vector machine model is insufficient for a multi-class task. To observe which method might yield better results, both approaches are considered.

Ensuring that the Support Vector Machine Classifier (SVC) achieves optimal performance requires that the model’s hyper parameters, namely \( C \), \( \gamma \), and the kernel, are chosen to yield the optimal model performance. Briefly, \( C \) specifies the scale of regularization. Thus, a small \( C \) allows for more regularization and vice versa. The kernel dictates the shape of the decision boundary, whether that be linear, polynomial, or the radial basis function. \( \gamma \) pertains particularly to the radial basis function, and it dictates the influence of individual training samples.

To identify these parameters, combinations of parameters are passed to the model and tested through the Exhaustive Grid Search method from SKlearn (Pedregosa et al., 2011). Exhaustive Grid Search iteratively tests various
combinations of parameters to identify which combination allows for the model to perform best for a given classification task. Following this, the optimal parameters identified were $C = 10$, a linear kernel, and, since the radial basis function (rbf) kernel was not utilized, the $\gamma$ parameter is not specified.

Stochastic Gradient Descent (SGD) is not on its own a unique classifier. Rather, SGD is an optimization method for unconstrained optimization problems. SGD is particularly effective when applied to problems with sparse input features and multiple dimensions. This makes the optimization method particularly applicable for this sentiment task. In implementation for this project, SGD is utilized with a linear kernel SVM as the classifier that the Gradient Descent algorithm optimizes in training.

4.1.3 Random Forest

The Random Forests algorithm is an ensemble algorithm, meaning it is composed of a multitude of Decision Trees from random samples drawn with replacement from the training set. The algorithm is an improvement upon the Decision Trees algorithm, where random creation of Decision Trees in the forest combats the tendency of high variance and over fitting in the Decision Trees method. Each tree within the Random Forest outputs a given probabilistic prediction, then the final classification is identified through averaging the probabilities and taking the resulting max as the prediction.

4.2 Model Evaluation

To ensure each model is evaluated on an uniform basis, a train and test split of the data is made initially, where 80% of the data is designated for training while 20% of the data is designated for testing. This allows each model a data set of roughly 1500 rows to be used for model testing following the training and validation phase for each model, as seen in Figure 1 below. Thus, the same training data set is passed to each model, along with the same TF-IDF matrix. The training data set is then iteratively split through K-Fold cross validation. For sake of uniformity in testing, each K-Fold instance is run $k = 10$ times, and the same random state is uniformly specified to ensure that the training and validation indices split are identical across each new instance of K-Fold, for each model being tested. So, each iteration of the K-Fold cross-validation trains $k - 1 = 9$ “folds” of the data in each iteration, then tested against 1 fold of the dataset. The logic of K-Fold can be seen best in Figure 2 below.
Figure 1: Initial train and test split logic.

Figure 2: Example of the K-Fold logic where k=4.

Each model is trained then tested with K-Fold cross validation, and evaluation metrics around model accuracy and the f score of each class are recorded. The accuracy score is defined to be the percentage of predicted labels from the model that match the actual labels from the validation data set. The formula for this statistic is:

\[
\text{acc}(c, n) = \frac{c}{n}
\]

where

\[
c = \text{total number of correctly labeled documents, and}
\]

\[
n = \text{total number of documents in the test set.}
\]
The f score is a measure of model accuracy derived from the precision, \( p \), and recall, \( r \), of each test. The precision of the model is its ability to make relevant classifications, while the recall is the proportion of relevant results that were correctly classified by the model. The classes of the data set are defined to be \( c_i \in \{C := c_1, c_2, \ldots c_n\} \) where \( i = 1, 2, \ldots n \) and \( n \) is the number of classes in the data set. Thus, the f score for a given class, \( f_{c_i}(p_i, r_i) \), is calculated as follows:

\[
    p_i = \frac{T_p}{(T_p + F_p)} \quad \text{(20)}
\]

\[
    r_i = \frac{T_p}{(T_p + F_n)} \quad \text{(21)}
\]

\[
    f_{c_i}(p_i, r_i) = 2 \cdot \frac{p_i \cdot r_i}{p_i + r_i} \quad \text{(22)}
\]

where each variable is defined to be:

\[
    T_p = \text{number of true positives,} \quad \text{(23)}
\]

\[
    F_p = \text{number of false positives,} \quad \text{(24)}
\]

\[
    F_n = \text{number of false negatives} \quad \text{(25)}
\]

Each iteration of K-Fold generates unique values for each performance metric. After iteratively running for \( k = 10 \) folds of K-Fold, the performance metrics are evaluated and stored in tabular format to compare each model’s performance relatively. Each performance metric is calculated with respect to each model, \( M_j \) for \( j := 1, 2, \ldots m \) where \( m \) is the number of models measured. So, the final metrics are calculated for \( k = 10 \) iterations of K-Fold as followed:

\[
    \text{Accuracy}(M_j) = \frac{\sum_{i=1}^{k} acc_i(c, n)}{k} \quad \text{(26)}
\]

\[
    \text{F Score (Positive)}(M_j) = \frac{\sum_{i=1}^{k} f_{c_1}(p_1, r_1)}{k} \quad \text{(27)}
\]

\[
    \text{F Score (Neutral)}(M_j) = \frac{\sum_{i=1}^{k} f_{c_2}(p_2, r_2)}{k} \quad \text{(28)}
\]

\[
    \text{F Score (Negative)}(M_j) = \frac{\sum_{i=1}^{k} f_{c_3}(p_3, r_3)}{k} \quad \text{(29)}
\]

The tests for each model are run \( k = 10 \) times per model, and the calculated results are listed in the table below.
### Table 1: Classification Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F Score (Positive)</th>
<th>F Score (Neutral)</th>
<th>F Score (Negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Naive Bayes</td>
<td>64.036</td>
<td>62.419</td>
<td>70.753</td>
<td>61.227</td>
</tr>
<tr>
<td>Multinomial Naive Bayes</td>
<td>73.190</td>
<td>75.909</td>
<td>75.672</td>
<td>75.235</td>
</tr>
<tr>
<td>Bernoulli Naive Bayes</td>
<td>70.495</td>
<td>69.939</td>
<td>74.393</td>
<td>72.311</td>
</tr>
<tr>
<td>Stochastic Gradient Descent</td>
<td>82.781</td>
<td>84.372</td>
<td>85.116</td>
<td>83.444</td>
</tr>
<tr>
<td>Random Forest</td>
<td>76.735</td>
<td>75.193</td>
<td>78.949</td>
<td>76.491</td>
</tr>
<tr>
<td>Linear Support Vector Machine (OVO)</td>
<td>84.690</td>
<td>84.926</td>
<td>86.288</td>
<td>82.617</td>
</tr>
<tr>
<td>Linear Support Vector Machine (OVA)</td>
<td>82.581</td>
<td>82.610</td>
<td>84.858</td>
<td>81.649</td>
</tr>
</tbody>
</table>

The OVO implementation performs notably better than the OVA model. I observe a $-2.109$ decrease in accuracy score for the OVA model compared to the LSVM OVO implementation. I can also observe a similarly significant decrease of the F scores for each class in the OVA case, with an observed $-2.316$, $-1.43$, and $-0.968$ percentage point loss for the positive, neutral, and negative classes respectively. Given the definition of the F score as a harmonic mean of the precision and recall of the classifier, improvement of the F score implies that the One-Versus-All implementation has not improved precision and recall. The OVO binary ensemble classifier thus yields fewer false positives and false negatives when compared to the OVA implementation and other machine learning algorithms. The OVA implementation does not necessarily improve the classification accuracy, while there are notable improvements in classification ability with the OVO implementation. There is a class imbalance when considering the OVA approach to multi-class classification. This class imbalance is likely a contributing factor to why the OVO implementation performs better when compared to the OVA implementation, as the imbalance can bias the model to predict the more common class, and is better seen in the BERT section below.

Overall, the performance of these classifiers are notable, but I can also explore one of the latest breakthrough models in text classification: BERT.

## 5 BERT

### 5.1 Background

The Bidirectional Encoder Representations from Transformers model, or BERT for short, is the application of Transformer architecture to machine learning for language modeling. BERT has models pre-trained on 2,500,000,000 English words from Wikipedia, and 800,000,000 words from the BooksCorpus, a corpus of English words from various pieces of literature (Devlin et al., 2018). Thus, in applications of BERT to text classification tasks, I am able to take advantage of transfer learning to improve the ability of the classifier in identifying word-to-word relations that can contribute to a given class (Peng et al., 2019).
A notable advantage to using a pre-trained BERT model, is the ability to fine-tune the neural network for a given classification task (Sun et al., 2019). Similar to other machine learning methods, such as support vector machines, a training and test data set can be used to fine-tune BERT for prediction. Sun et al. (2019) showed that this can be done for multi-class text classification tasks, and can achieve notable classifier performance even with a small amount of training data.

In application, there are two pre-trained neural networks available for use: BERT\textsubscript{base} and BERT\textsubscript{large}. The difference between these models lies in the size of the architecture on which they are built. BERT\textsubscript{base} has fewer layers and heads, resulting in a network of about 110,000,000 parameters, while BERT\textsubscript{large} has more layers and heads than the base model, amounting to 340,000,000 parameters in the network.

Literature shows that BERT is a leading methodology for text classification (Devlin et al., 2018; Li et al., 2019; Sun et al., 2019). Thus, the BERT model is utilized for this paper. For the faster training times due to decreased number of parameters, while maintaining notable classification performance, BERT\textsubscript{base} is used. Furthermore, in the implementation of BERT in this study, parameter recommendations are followed as per Devlin et al., 2018.

To utilize BERT in this study, the pretrained model BERT\textsubscript{base} from the transformers (Wolf et al., 2020) library is utilized for sequence classification. For the tokenization of raw text data for use in the BERT model, the BertTokenizer is also utilized from the transformers library. Each row of text in the train and test set are passed through the tokenizer, then to the BERT classifier. The classifier then outputs the probability a given input is associated with each class, through the use of tensorflow’s softmax function. These probabilities are finally evaluated through conditionals to identify the final label for the input.

5.2 Preparation

Given that the problem is a multi-class classification task, One-Versus-One and One-Versus-All approaches in the context of BERT are both considered and evaluated. For these structures, the training data set is replicated and divided, and can be seen again in the table below.
<table>
<thead>
<tr>
<th>Method</th>
<th>Classes</th>
<th>Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>positive vs. neutral</td>
<td>5,172</td>
</tr>
<tr>
<td>OVO</td>
<td>positive vs. negative</td>
<td>4,984</td>
</tr>
<tr>
<td></td>
<td>neutral vs. negative</td>
<td>4,984</td>
</tr>
<tr>
<td></td>
<td>positive vs. {neutral or negative}</td>
<td>7,570</td>
</tr>
<tr>
<td>OVA</td>
<td>neutral vs. {positive or negative}</td>
<td>7,570</td>
</tr>
<tr>
<td></td>
<td>negative vs. {positive or neutral}</td>
<td>7,570</td>
</tr>
</tbody>
</table>

To further understand this methodology, I can first consider the example of the positive versus neutral model training data set for the OVO method. Here, a subset of the main data set is taken where only rows with the pre-labeled positive and neutral classes are kept in the training data set, implying the 2,398 rows of negative classes are removed from this subset. When the model is fine-tuned with this data set, the output classes considered will be positive or neutral alone. I may also consider the positive versus {neutral or negative} model training data set used for the OVA approach. Here, the main data set is first replicated. Then, each class label for each row is considered. Where a row label is positive, the label is kept. When a neutral or negative label is encountered, the label is then changed to Not Positive. The purpose of this is to allow the model to still consider all available training data, while only classifying positive tweets. Thus, for this example, the model training data set is then composed of 2,598 positive class rows and 4,972 Not Positive class rows.

### 5.3 Binary Ensemble Model

In some cases, a single classifier can be used with reasonable levels of accuracy for a multi-class classification problem, namely with the classes: positive, neutral, negative. Yet using a single classifier means that the model is only able to be trained on and predict the probability of a document being either positive, neutral, or negative. While this approach can be implemented with reasonable levels of accuracy, more information can be derived from a training data set to improve the accuracy of prediction for a model by focusing on both the probability of a document being a class and the probability of a document not being a class. So each class is then considered on an individual basis rather than collectively in a single model. Instead of having one classifier predicting multiple classes, an ensemble of binary classifiers is utilized, where each classifier is trained to identify if a document is a class or not a class.

### 5.4 Training Data for Binary Ensemble

In the case of sentiment analysis for this study, a classifier is trained for each class: positive, neutral, and negative. In this paper, I implemented and evaluated a One-Versus-One (OVO) approach and One-Versus-All (OVA) of SVMs above, and now will examine these same implementations but for BERT. So, in addition to the original classes out-
lined above, I introduce three new classes in the OVA method: *not positive*, *not neutral*, or *not negative*. These classes are introduced to the model by re-examining the training data set.

For the three classes considered in this study, the training data set is duplicated three times and organized by *positive*, *neutral*, and *negative* classes. Each training data set is then examined by class. So, for the *positive* class, every document in the *positive* training data set is examined. If the label of the document is recorded as *positive*, then the label is kept as is. For any document encountered that is labeled as *neutral* or *negative*, the document is labeled instead as *not positive*. This leaves the training data set with 2,586 *positive* documents, and 4,984 *not positive* documents. This process is repeated for the *neutral* and *negative* training data sets respectively. The data sets for the OVA implementation as well as the OVO implementation and their respective class balances are visualized in the table below.

<table>
<thead>
<tr>
<th>Method</th>
<th>Classes</th>
<th>Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVO</td>
<td>positive vs. neutral</td>
<td>5,172</td>
</tr>
<tr>
<td></td>
<td>positive vs. negative</td>
<td>4,984</td>
</tr>
<tr>
<td></td>
<td>neutral vs. negative</td>
<td>4,984</td>
</tr>
<tr>
<td>OVA</td>
<td>positive vs. {neutral or negative}</td>
<td>7,570</td>
</tr>
<tr>
<td></td>
<td>neutral vs. {positive or negative}</td>
<td>7,570</td>
</tr>
<tr>
<td></td>
<td>negative vs. {positive or neutral}</td>
<td>7,570</td>
</tr>
</tbody>
</table>

While the case of the OVO implementation has generally equal numbers of classes, the OVA implementation does not necessarily see the same class balance. In the case of a OVA implementation, a class imbalance can be seen between each of the two classes considered. This means that, when training each individual model in the ensemble, the training data has more rows of data for one class relative to the other. In the case of the training data set for the *positive* model, there are 2,586 rows for the *positive* class, yet 4,984 rows for the *not positive* class; nearly 2 times as many *not positive* rows relative to *positive* rows. When understanding this problem and how it has been addressed in the literature, Rifkin and Klautau, 2004 points out how despite the perceived shortcoming of this class imbalance, OVA implementations can still perform admirably relative to OVO approaches. The theory behind the threat of this class imbalance is that the classifier will then be biased to classifying the majority class. Yet when considering class imbalance, BERT itself is more resilient toward the problem when compared to a classifier such as SVMs (Madabushi et al., 2020).

Once each training data set is relabeled for the respective classes, a BERT model is trained on each respective data set. This creates three binary classifiers, with each classifier being training to identify *class* or *not class*. The
binary classifiers are then organized into an ensemble via a series of conditional statements. Each conditional statement decides what label to assign to a given document when passed through the ensemble, based on the outputs of each classifier within the ensemble. Each classifier then, when predicting a label, outputs the probability that the given document is or is not the class specified by each respective model. The algorithm logic for OVA can be seen in Figure 3.

```python
1 def binary_ensemble(df_row):
2     pass text to classifiers and get dictionary of outputs from each
3     get max probability from each model
4     count the number of class probability versus not class probability
5     if number of not class probability == 2:
6         output label = class
7     elif number of class > 1:
8         output label = max(probability class)
9     elif number of not class == 3:
10        output label = min(not class)
```

**Figure 3: Binary Ensemble Logic**

What is advantageous about this approach is being able to consider the edge case of when none of the three binary classifiers return a *positive*, *neutral*, or *negative* class label. In the case of Figure 3, I recognize this consideration in line 19, where I consider when each of the three probability class labels are "not class.” In these instances, each classifier outputs a series of probabilities which are interpreted as each classifier’s confidence in a document being labeled a given class.

So generally following this logic, I am able to pick between the three classifiers through identifying the maximum. Various sample outputs can be seen below, and where the bold item in each row is the chosen output label of the ensemble.
<table>
<thead>
<tr>
<th>Positive Binary Classifier</th>
<th>Neutral Binary Classifier</th>
<th>Negative Binary Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Label</strong></td>
<td><strong>Probability</strong></td>
<td><strong>Label</strong></td>
</tr>
<tr>
<td>positive</td>
<td>0.9899922</td>
<td>not_neutral</td>
</tr>
<tr>
<td>positive</td>
<td>0.8876222</td>
<td>neutral</td>
</tr>
<tr>
<td>not_positive</td>
<td>0.6843332</td>
<td>not_neutral</td>
</tr>
<tr>
<td>positive</td>
<td>0.7886455</td>
<td>neutral</td>
</tr>
</tbody>
</table>

Table 4: Example probability values from an OVA ensemble

While Table 4 represents examples for the OVA approach, similar logic is followed in the OVO approach. Given that each model in an OVO ensemble only predicts between two classes, I may follow a majority voting technique. So, various examples outputs are explored below and bold entries are highlighted as the final outputs of the classifier.

<table>
<thead>
<tr>
<th>Positive vs. Neutral Classifier</th>
<th>Positive vs. Negative Classifier</th>
<th>Neutral vs. Negative Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Label</strong></td>
<td><strong>Probability</strong></td>
<td><strong>Label</strong></td>
</tr>
<tr>
<td>positive</td>
<td>0.9899922</td>
<td>positive</td>
</tr>
<tr>
<td>neutral</td>
<td>0.8876222</td>
<td>negative</td>
</tr>
<tr>
<td>positive</td>
<td>0.6843332</td>
<td>negative</td>
</tr>
<tr>
<td>neutral</td>
<td></td>
<td>neutral</td>
</tr>
</tbody>
</table>

Table 5: Example probability values from an OVO ensemble

The first two rows of Table 5 showcase the majority voting technique, where a given class is identified by two of the binary models as the likely label for the given document, and thus output as the final label. In the third row, I see an instance where the maximum logic from the OVA method is then utilized, as the majority vote technique is non-applicable.

### 5.4.1 BERT OVO vs. OVA Evaluation

Once each model is trained, the BERT models are organized into an ensemble. Each classification the BERT model processes is able to output the predicted class with probability scores, similar to a support vector machine based model. So with both the OVO and OVA implementations, the final output class is determined by the maximum logic.

Having each of the OVO and OVA approaches implemented into an ensemble, the final classifiers are then evaluated against the validation set. This data set is composed of 1514 rows, with an even class balance of about 500.

WILLIAM & MARY

BLOCKCHAIN IN HEALTHCARE: A NEW PERSPECTIVE FROM SOCIAL MEDIA DATA
rows per class. To ensure that the performance metrics are accurate and follow a uniform methodology relative to the previous models, K-Fold is also implemented in this case with \( k = 10 \) folds. The performance metrics obtained from the K-Fold iterations are reflected in the table below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F Score (Positive)</th>
<th>F Score (Neutral)</th>
<th>F Score (Negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Naive Bayes</td>
<td>64.036</td>
<td>62.419</td>
<td>70.753</td>
<td>61.227</td>
</tr>
<tr>
<td>Multinomial Naive Bayes</td>
<td>73.190</td>
<td>75.909</td>
<td>75.672</td>
<td>75.235</td>
</tr>
<tr>
<td>Bernoulli Naive Bayes</td>
<td>70.495</td>
<td>69.939</td>
<td>74.393</td>
<td>72.311</td>
</tr>
<tr>
<td>Stochastic Gradient Descent</td>
<td>82.781</td>
<td>84.372</td>
<td>85.116</td>
<td>83.444</td>
</tr>
<tr>
<td>Random Forest</td>
<td>76.735</td>
<td>75.193</td>
<td>78.949</td>
<td>76.491</td>
</tr>
<tr>
<td>Linear Support Vector Machine (OVO)</td>
<td>84.690</td>
<td>84.926</td>
<td>86.288</td>
<td>82.617</td>
</tr>
<tr>
<td>Linear Support Vector Machine (OVA)</td>
<td>82.581</td>
<td>82.610</td>
<td>84.858</td>
<td>81.649</td>
</tr>
<tr>
<td>BERT - OVO Ensemble</td>
<td>94.900</td>
<td>94.238</td>
<td>94.945</td>
<td>95.577</td>
</tr>
<tr>
<td>BERT - OVA Ensemble</td>
<td>91.549</td>
<td>90.826</td>
<td>92.413</td>
<td>91.604</td>
</tr>
</tbody>
</table>

A notable performance increase is observed by utilizing BERT neural networks in place of support vector machines in the ensemble implementation. When compared to the binary ensemble OVO and OVA SVM implementations, I see notable accuracy score improvements of \(+10.210\) percentage point increase for the BERT OVO ensemble, and a \(+8.968\) percentage point increase for the BERT OVA ensemble. I also see that given the F score for each respective class, I see notable improvements with the BERT OVO ensemble outperforming the OVA ensemble.

### 5.5 Model Application

From the exploration of various machine learning methods, BERT in a One-Versus-One Ensemble has been identified as the most accurate against the training data set. Thus, I am able to use this model and apply it to the greater data set of 207,287 tweets on blockchain and healthcare. Maximum logic is applied, and final sentiment classification labels are generated for each tweet in the data set. The results of this classification are then explored.

### 6 Results

#### 6.1 General

The raw data set acquired through the Twitter API consists of 207,287 rows. Each of these rows corresponds to a unique tweet retrieved from the keyword search of “blockchain healthcare,” as in the section above. Each unique
tweet has specific data associated with it, as retrieved from Twitter. This data includes time-specific information associated with both the time of posting of the tweet, the username and screen name of the account posting said tweet, and the raw text of the tweet. The raw text is pre-processed by the methodology mentioned in section 3.2, and classified by the BERT ensemble. The final label of each tweet is then determined as positive, neutral, or negative, and is considered the sentiment of the tweet.

6.1.1 Bots

When looking at Twitter data, it is largely known that accounts exist on social media platforms that are programmaticaloy operated through software implementations, known as bots (Bessi and Ferrara, 2016, Heredia et al., 2018, Stella et al., 2018). These accounts, sometimes posing as people or organizations, are able to tweet repetitively about a topic, and can potentially alter the true sentiment observed about a topic. For example, an organization could use bot accounts on Twitter to tweet out positive viewpoints about a given topic, should it benefit the organization. This would then artificially inflate the positive class.

Seeing the implications bot accounts can carry toward understanding sentiment on Twitter, a known classifier in the space of Twitter research, dubbed Botometer (Sayyadiharikandeh et al., 2020), is utilized to classify bot accounts in the data set. Thus, each user account encountered in the Twitter data set was passed to Botometer, and a probability score is returned indicating the confidence of the model in classifying an account as a bot. Following the documentation, a probability score of above 95% was then labeled a bot account, while any score less than this threshold were indicated as a not bot account. One noteable limitation to this is that some Twitter accounts could not be classified due to being either private, restricted, or no longer existing. As such, there was an "N/A" classification for these accounts.
In Figure 4, I can immediately see that over 95% of the accounts in the data set are classified as not being bot accounts. So, only 1.47% of accounts, or 661 accounts, are identified as bots in the data set. It is worth noting that the model could not access 3.66% of accounts, or 1,642 accounts. Thus, these accounts are excluded from the data set, as they cannot be determined as being or not being bot accounts. For further analysis, bot accounts are filtered out from the data set. This accounts for the removal of 8,479 tweets. This is to ensure that the sentiment observed further in this analysis is not construed by posts from these bot accounts.

6.2 Volume

In Figures 5a and 5b, the number of tweets are mapped in a histogram relative to the date on which the tweet was posted. The volume of the tweet data is visually observed to follow normal distribution centered around the year 2018. This entails an observable peak in tweet volume in the year of 2018. This peak can be further quantified as a net 25,886 tweet increase from the previous year, where 37.65% of the total tweets from 2016 to early 2021 are accounted for in 2018. Further aggregating to months, April and May of 2018 are seen as the points of peak tweet volume.
Figure 5: (a) The number of tweets by post date (year) showing percent of total as well as the total number of tweets. (b) The number of tweets by post date (month) with a 12-month moving average indicated by the orange line.

More than a third of the total tweets in the data set occurring in 2018, as seen in Figure 5(a/b), is noteable. To understand why this may have happened, it is necessary to understand the background of blockchain. Blockchain technology was theoretically established in the 1980s through a dissertation (Sherman et al., 2019). Yet one of the first production applications of the technology was in the founding of the cryptocurrency Bitcoin in 2008 (Sherman et al., 2019). Given the cryptocurrency’s status as one of the first real applications of blockchain, Bitcoin’s value appreciation over 2017 likely brought the public and professional eye toward the currency. Bitcoin as a security started 2017 achieving a price value of $1000 in January, the first time reaching that milestone in the previous three years. Bitcoin ended 2017 worth $20000 a coin. This massive increase in value of Bitcoin as a security likely brought more attention not only to the cryptocurrency but to what the cryptocurrency is built upon: blockchain. This hypothesis is likely a contributing factor to what caused the increased public reference to blockchain following Bitcoin’s value rise in 2017.

Drawing this attention to Bitcoin inherently drew industry’s eye to blockchain. Bitcoin as one of the first real applications of blockchain technology proved it could provide real value in implementation. Therefore, 2018 is understood to be the third phase on the life cycle of blockchain: implementation (Frost Sullivan, 2019). Early the

1https://www.coindesk.com/price/bitcoin
following year, governments such as in Switzerland began accepting Bitcoin for tax payments\textsuperscript{4}, and by November of 2018 venture capital investment into blockchain-based technology surpassed $1 billion dollars for the first time and reached to nearly $4 billion dollars\textsuperscript{5}, while over 75\% of banks reported exploration into blockchain technology the following month\textsuperscript{6}. This shift in investment toward companies utilizing blockchain technology signified the beginning of the period of implementation, where players in a variety of sectors emerged to bring the theorized business value of blockchain to industry leaders as well as to emerge as key players through utilization of the technology for their own venture. Thus, seeing the tweet volume spike in 2018 at the start of the implementation phase of blockchain is understandable with the provided context.

\textsuperscript{4}https://www.ccn.com/swiss-town-will-accept-bitcoin-tax-payments-2018/
\textsuperscript{5}https://www.coindesk.com/vc-investment-in-blockchain-startups-is-up-280-so-far-this-year
6.3 Sentiment

Beyond tweet volume, I will take a closer look at the classification labels: *positive*, *neutral*, and *negative*, of the tweets in the data set. In the figure below, I am able to see the distribution of the classes in the entire data set.

![Figure 6: Percent distribution of classes in the data set.](image)

From Figure 6, I can see the *positive* class is observed to be the most frequent class in the data set, accounting for 49.80% of the tweets, or 99,014 tweets. The *negative* class is found to have the lowest rate of occurrence, with only 8.45% of tweets, or 16,797 tweets, identified as such. Thus, there are 82,217 more *positive* tweets than *negative* tweets, amounting to a ratio of roughly 6 *positive* tweets for each *negative*. This vast difference favoring *positive* tweets indicates that the general public perception of blockchain in healthcare, from the Twitter perspective, is *positive*.

While it is important to observe *neutral* tweets so as they are not classified as either *positive* or *negative*, in terms of analyzing the public perception of tweets over time, it is less informative to focus on *neutral* class tweets. The function of identifying these tweets assists in the classifier’s identification of tweets that do not necessarily indicate a polar opinion of a user, such as a neutral news headline or a company announcing an upcoming webinar. This still allows us important information especially around the volume of tweeting observed in the figures above, but is less applicable when understanding the public sentiment of blockchain in healthcare. Thus, it is observable that public sentiment is heavily skewed *positive* in the data set relative to the *negative* class tweets, and this is indicated by the 49.80% of tweets being labeled as *positive* by the classifier.
I can examine further how the sentiment changes over time. It is observable in Figure 7 that the earlier months of the data set showcase largely neutral sentiment. As time increases, positive visually appears to compose a greater percentage of the monthly class distributions. That being said, negativity is still apparent. It is difficult to visually identify a quantified trend in the negative class composition over time. This notable increase in the positive class tweets beginning in 2018 can be further identified in the box plots below for each class.

Figure 8: Box plots for Positive class over Year

Figure 9: Box plots for Positive class grouped before versus after 2018
**Figure 10:** Box plots for Neutral class over Year

**Figure 11:** Box plots for Neutral class grouped before versus after 2018

**Figure 12:** Box plots for Negative class over Year

**Figure 13:** Box plots for Negative class grouped before versus after 2018
Figure 14: Percent distribution of classes broken down over time aggregated by month.

To quantify this, the data can be aggregated on a timeline from 2016 to early 2021 by months, and a linear line of best-fit is applied to each class. The formulae for each line of best-fit can be seen in the Table 4 below.

<table>
<thead>
<tr>
<th>Class</th>
<th>Formula</th>
<th>p-value ($\alpha = 0.05$)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>$%$ of class tweets $= 2.01257 \cdot 10^{-5} \cdot Post.Date - 0.782013$</td>
<td>0.0386471</td>
<td>0.0704855</td>
</tr>
<tr>
<td>Neutral</td>
<td>$%$ of class tweets $= -1.99422 \cdot 10^{-4} \cdot Post.Date + 9.06817$</td>
<td>$&lt; 0.0001$</td>
<td>0.49406</td>
</tr>
<tr>
<td>Positive</td>
<td>$%$ of class tweets $= 1.59081 \cdot 10^{-4} \cdot Post.Date - 6.40503$</td>
<td>$&lt; 0.0001$</td>
<td>0.4673846</td>
</tr>
</tbody>
</table>

Table 6: Lines of best-fit by sentiment class.

For each line of best fit, observed p-values are sufficiently less than an $\alpha = 0.05$, so the variable Post.Date, which is represented by months in this case, is statistically significant. While it is worth noting that the $R^2$ values are low, and thus not necessarily indicative of Post.Date being the most explanatory variable of tweet volume over time on its own, that does not discount that the Post.Date statistically affects the tweet volume, in the case of each class, over time. As time increases, the tweet volume of negative tweets increases at a rate of $2.01257 \cdot 10^{-5}$ percentage points. While, in the case of positive classes, the tweet volume increase at a rate of $1.59081 \cdot 10^{-4}$ percentage points as time increases, a notably greater rate of increase when compared to the negative class tweets. The positive and negative tweets compose a greater share of the tweets in a given month over time. The opposite is observed of neutral tweets.
tweets over time. In the case of the neutral class, the percentage point composition in a given month decreases by $-1.99422 \cdot 10^{-4}$ percentage points as time increases. The polarity of tweets increases as time increases, yet positive tweets compose a greater percentage of the tweets in a given month as time increases, when compared to negative tweets. In the public realm of Twitter, users tweeting in the blockchain and healthcare sphere are gradually more positive about the technology when associated with healthcare, yet the increase in negative tweets shows that users are still maintaining skepticism about blockchain.

Seeing the rate of positive tweets increasing at a greater rate than negative tweets indicates that sentiment toward blockchain in healthcare is increasingly more optimistic over time. This implies that accounts tweeting about this topic over time have become more optimistic than pessimistic toward the technology’s potential in the healthcare sector. This trend in sentiment coincides with blockchain’s development from the identification of theoretical applications to the realization of the technology’s value through adoption by current players in the healthcare system and through the entrance of new players as well (Frost Sullivan, 2019). Companies have gradually been testing the technology for implementation at the production level. This can vary from applications in patient data management, to regulatory compliance, to improving current systems such as Contracts and Chargebacks for pharmaceutical companies (Frost Sullivan, 2019). Thus, given the wide range of applications being explored, the sentiment observed indicates that public perception maintains optimism through the proof-of-concepts and various other pilots of blockchain as a solution.

Thus far, it is observed that the volume of the tweet data visually mimics a bell curve centered around the year 2018. Furthermore, as time increases, polarity of tweets increases. Of note is that the rate of increase of positive tweets is notably greater than the rate of increase of negative tweets. Understanding the overall trend within the data, company versus personal accounts are aggregated to understand how their respective sentiments are observed.

### 6.4 Account Type Classification

An important viewpoint to understand how blockchain is viewed in the healthcare sector is to consider the kinds of users tweeting opinions about the technology. For this, the two categories of users can be identified as “company” accounts or “personal” accounts. A company account is defined to be a user account associated with a business entity, while a personal account is defined to be a non-organization account.

To programmatically make these classifications, a general purpose classifier is utilized called m3inference (Wang et al., 2019). This classifier is a neural network that can classify information about a Twitter user account such as age of the user, gender of the user, and if the account is an organization or a personal account. For applications in this study, only the organization status of the account is utilized. Thus, for application, the various necessary parameters
for classification of each Twitter account are passed to the model, and a classification of “company” is made for organization accounts, and “personal” is made for personal accounts. I can finally aggregate the user accounts in the data set by their status.

Figure 15: Percent of company versus personal accounts.

Figure 16: Total tweets of company versus personal accounts.

Figure 15 shows that the majority of the users involved in the data set are personal accounts, accounting for 63.89% of users or 27,252 users. Yet to further understand the account types’ contributions to the data set, Figure 16 is made. Thus, it is seen that not only do personal Twitter accounts make up the majority of accounts in the data set, but the personal accounts also produce the majority of tweets in the data set. Specifically, personal accounts produce 57.07% of the tweet data, or 113,456 tweets. This entails that each personal account tweets roughly 4.16 tweets per capita. Yet, with the same context, company accounts tweet at a rate of roughly 5.52 tweets per capita. So, on a per-account basis, company accounts tweet at a higher rate than personal accounts.

Companies are likely tweeting more frequently than personal accounts because of how they utilize Twitter as a platform. A company could be promoting itself through tweets about webinars or news headlines to name a few examples. Personal accounts, on the other hand, could tweet in reaction to what companies are saying. Yet a personal account by enlarge will not tweet solely in the blockchain and healthcare sphere. On the other hand, a company account can represent an entity with a vested interest in the blockchain and healthcare space. Whether it be a blockchain start up or a major company involved in the healthcare sphere, these companies will logically tweet more consistently about blockchain healthcare. Thus, company accounts tweeting more frequently exemplifies their heightened involvement in discussing blockchain and healthcare, while personal accounts tweet less frequently in the sphere but occur in a greater number of unique accounts. Understanding the rate of tweeting by these two groups of accounts, I can now understand with what distribution of sentiment these accounts tweet.
Figure 17: Sentiment class distribution of total tweets aggregated by account type.

As seen in Figure 17, there is a notable difference between the frequency of positive and neutral tweets produced by companies relative to personal accounts. More specifically, company accounts see a 2 percentage point difference between positive and neutral tweets, while personal accounts see a more substantial 12.61 percentage point difference for the same classes. This intuitively makes sense, because a company tweeting about blockchain and healthcare is more likely to tweet about company related events or news. A company could tweet about a webinar they are holding or recent news about a blockchain and healthcare company partnership, which would be marked as neutral. On the other hand, a personal account tweeting about the topic will be more likely to comment opinions on an event. A company could tweet about a new blockchain and healthcare partnership they are entering, a neutral tweet, yet a personal account could see this tweet and comment on it with a more polar response (either positive or negative).

When compared to the initial hypothesis about companies tweeting more positively due to vested interest, these results are particularly of note. Companies tweeting more neutrally, relative to positive tweets, implies that this vested interest may not necessarily be manifested. One reason this may be occurring is again due to the nature of company tweets as highlighted above. Companies use the Twitter platform to promote their business while also announce news or business-related events. So, when identifying positive sentiment among company tweets, the overall distribution of tweets accommodates for the dilution of positive and negative classes created by this nature of tweeting. Personal accounts, on the other hand, do not necessarily deal with this same problem. Personal accounts do not represent a brand or company, and thus would not use the Twitter platform to announce news. Rather, these accounts have the flexibility to express personal reactions to various events. So, while a company account may tweet about
a new webinar they are hosting, a personal account tweet about this webinar and express more polarized emotion about the topics. These differences in how companies utilize Twitter when compared to personal people is a noted challenge of using Twitter data to understand this problem. One way to accommodate for this is to examine only the polar classes and ignore neutral tweets.

Figure 18: Sentiment class distribution of total tweets aggregated by account type and excluding the neutral class.

By excluding the neutral tweets, I see in Figure 18 that both companies and personal accounts tweet overwhelmingly positive relative to negative. Yet, when comparing between account types, I again do not observe the vested interest outlined earlier in the form of more positive tweets coming from companies relative to personal accounts. From Figure 18, I even go as far as seeing personal accounts having a class distribution of 2.22 percentage points more positive tweets when compared to company accounts. So, not only do I not visually observe companies’ vested interest inflating sentiment relative to personal accounts, I even observe personal accounts tweeting slightly more positively when compared to companies.

Even when accounting for the notable difference in rate of neutral tweets, both the company and personal accounts tweet overall positive more frequently than negative. This happens despite the expectation that a company would tweet more positively about this technology as it is emerging in the healthcare industry, relative to non-companies as observers to the technology’s emergence having the ability to be more critical of blockchain in healthcare. Non-company participants in the discussion of blockchain in healthcare are maintaining optimism for the technology’s successful implementation in the sector and companies are understandably hopeful for the success of the technology as well.
7 Discussion

7.1 Post-hoc Work

To validate the findings on the social acceptance of blockchain technology in healthcare, I conducted post-hoc interviews with organizations in the healthcare sector. These interviews help us pinpoint the challenge in integration of the technology for this industry. Factors such as poor scalability, high cost of implementation, and challenges associated with the performance of the technology are hurdles still needed to be overcome through implementation of blockchain (Chukwu and Garg, 2020). While these challenges are observed, there is yet to be a concrete example of a proof-of-concept proving blockchains business value add in a production level setting. Yet that is not to say that companies are not exploring the technology.

An example of blockchain implementation can be found in the industry response to the Drug Supply Chain and Securities Act (DSCSA). The DSCSA is legislation that was passed in 2016 to impose stricter regulations on the pharmaceutical supply chain to improve industry ability to identify and remove potentially dangerous or bad-batches of drugs from the supply chain. In early 2019, the Food and Drug Administration began pilot projects to explore solutions for the regulatory compliance within the supply chain, and blockchain was evaluated as a potential industry solution. The key players involved in these blockchain pilots have included major pharmaceutical industry stakeholders such as Merck, Pfizer, Bayer, and IBM. One notable company leading a pilot project was Mediledger. To understand how this implementation has actually gone for industry-players, I spoke with a representative from Mediledger.

I spoke to a key blockchain-based solution provider involved in the FDA pilot, Mediledger, which is an industry-led blockchain powered network in the life sciences industry and is run by Chronicled. According to the representative, Mediledger is still exploring different applications of the technology while a production-level realization of blockchain’s value has yet to be realized. In the context of the FDA pilot program, the blockchain provider sees blockchain as a potential solution; however, the nature of the application being for regulatory compliance means that the adoption is driven less by business value and more by requirement from the government to produce a solution. Thus, the actual application of the technology will take many years until the solution is needed. Yet despite potential hurdles in logistical adoption, blockchain is still a competitive solution. Mediledger’s pilot project report was released in early 2020, and outlined the success of the pilot project with partners. The report highlighted blockchain’s ability to effectively meet DSCSA guidelines while maintaining a low-error, high speed, and reasonable-cost im-

https://www.mediledger.com/
https://www.chronicled.com/about-us
plementation. While the pilot project was a success for blockchain, the hurdles mentioned by the Mediledger representative persisted in the report. Thus, an actual implementation of blockchain in the FDA pilot will take time, and even then it is not guaranteed that blockchain will be adopted as the solution of choice with costs of implementation and industry consensus considered.

To further understand how the piloting of blockchain as a regulatory compliance solution for the DSCSA is faring, I spoke with representatives from the Partnership for DSCSA Governance (PDG). This organization is a “forum dedicated to developing, advancing, and sustaining an effective and efficient model for interoperable tracing and verification of prescription pharmaceuticals in the US.” The organization formed as a governing body for ensuring industry compliance with standards set forth in the DSCSA regulations and serves as a forum for major players in the pharmaceutical supply chain industry. They are focused on defining the business processes and technical systems necessary for the pharmaceutical supply chain to adopt, with the further goal of overseeing the technical infrastructure when in place. They repeatedly convene industry throughout the pilot and discuss findings from various stages in the program. Despite cited successes of the pilot from the pharmaceutical supply chain stake holders, PDG has seen challenges arise. One example they mentioned was around technical challenges and how, even though blockchain has seen success in its pilots with smaller groups of the pharma supply chain, the hurdles presented by different players in the supply chain opting for different blockchain-based solutions challenges blockchain’s standing as the first choice technology for compliance. If some stakeholders adopt one blockchain network and other stakeholders adopt a different one, the ability for these networks to communicate is a major technical challenge. While PDG can govern a regulatory solution across the pharmaceutical supply chain, they cannot require any one solution be adopted by each stakeholder. Thus, the diversity of solutions presented and how various solutions, blockchain-based or not, can be adopted by industry makes interoperability a major hurdle for blockchain to be implemented at the production level. The PDG representatives further mentioned how, despite one’s potential belief that blockchain is an all or nothing solution, realistically a blockchain-based solution would satisfy only a piece of the regulatory compliance puzzle. If a solution is adopted as a piece of the puzzle, it needs to be accepted across all stakeholders in the supply chain. Thus, blockchain cannot only be adopted by a portion of supply chain.

These conversations with key players in the healthcare industry show how, even though blockchain has progressed through pilots and been proven to yield business value, it is not guaranteed that the technology will be applied. The FDA pilot project exploring blockchain is just one industry example of how blockchain still faces notable hurdles in implementation (Zou et al., 2020). The key points of business value blockchain can provide are contingent on participation across industry, yet another hurdle of industry consensus that blockchain would need to surmount.

---

10 https://www.mediledger.com/fda-dscsa-pilot-project-success
11 https://dscsagovernance.org/mission-vision-2/
Thus, despite the theoretical promise it has shown and attention it garnered from 2017 to 2018, the challenges of implementation manifest in 2019 and 2020 as public discourse expresses optimism in what blockchain is capable of, but the volume of discussion dampens as time progresses. Since there has yet to be a successful production-level implementation, the decrease in volume of tweets shows how the implementation phase continues on to the present day.

In the continuation of the implementation-phase of the technology, there is still strong optimism for the technology. From the case of the FDA pilot projects, I have seen multiple pilot projects conclude and report successful results toward blockchain’s feasibility as a potential solution for adoption (Chien et al., 2020)\textsuperscript{12, 13}. So this noted success in controlled proof-of-concept settings could be a contributing factor to the increasing positive sentiment over time. Industry’s progression with blockchain as a solution and consideration among some of the largest players in the pharmaceutical supply chain indicates optimism toward the technology’s potential application across industry. Yet, even with optimism, the public eye seems to be waiting for a successful implementation to be seen in the industry beyond pilots and theoretical proof-of-concepts.

7.2 Conclusion

In this study, I analyzed a data set spanning from January 2016 to January 2021, and from this I have seen a relatively stronger increase in the rate of occurrence of positive tweets as well as notable, yet not as stark, increase in negative tweets. A notable increase in the frequency of occurrence of positive tweets begins in 2018, and gradually grows albeit at a smaller rate. In terms of volume, I have seen the number of mentions of blockchain in healthcare make a drastic spike in 2018, yet taper off in the following years. Thus the data shows how, as the implementation phase of blockchain began in 2018 (Frost Sullivan, 2019), discussion of blockchain in healthcare increased alongside the sentiment of this discussion becoming more positive. This implementation was in investment into blockchain technology-based companies (Frost Sullivan, 2019)\textsuperscript{14} as well as key players exploring how blockchain could potentially benefit their organization. Yet despite public discussion being largely positive, the challenges of implementation of blockchain persisted. This can be seen as the volume of tweets around the topic gradually decrease following 2018 while negative class tweets increase at a slow rate. So, the increase in positive sentiment indicates that public discourse over blockchain in healthcare is hopeful for successful implementation; however, actually integrating the technology into the healthcare sector has been challenging.

From this methodical approach, a One-Versus-One ensemble of BERT models was found to be the most performant based on the training data curated for this task. Thus, when applied to the question of understanding the

\textsuperscript{12}https://www.ledgerinsights.com/ibm-merck-walmart-blockchain-fda-pharmaceutical-pilot/
\textsuperscript{13}https://www.mediledger.com/fda-dscsa-pilot-project-success
\textsuperscript{14}https://consensys.net/blog/news/the-decade-in-blockchain-2010-to-2020-in-review/
public perception of blockchain and healthcare, tweets were seen to be much more frequently *positive* rather than *negative*. When further understanding how sentiment varies between personal and company Twitter accounts, to address vested interests in the success of blockchain in healthcare, it was found that personal and company accounts produce *positive* tweets much more frequently than negative tweets. Thus, it is concluded that despite a company having a more vested interest in the success of the technology, both company and personal Twitter users discuss blockchain in healthcare in a positive manner. Further understanding of the inverse trend of tweet volume decreasing while *positive* sentiment increases over time motivated us to explore more directly the industry perspective through conversations with participants in a major exploration of blockchain’s application. From this, I found these stakeholders perspective of proof-of-concepts of blockchain to be successful yet still perceived to face many hurdles before production-level implementation can be achieved. Thus, optimism is maintained, yet the wait continues for a notable production-level application to be observed.

Future work could be done to improve the performance of the model, and its ability to better identify context within tweets. While Twitter is a good place to begin this research, more work could certainly be conducted in expanding the pool of data collected. This could include, but is not limited to, expanding the Twitter keyword search to include more combinations of terms, or looking at different mediums such as the comments section in various news articles or analyzing content published in relevant journals. In terms of understanding how industry perceives blockchain as a solution, more work could be done in conversing with industry stakeholders as well as looking more closely at various use-cases being explored today.
References


