Mitigating Risk in Financial Industry by Analyzing Social-Media with Machine Learning Technology

Issayas M. Haile and Yanzhen Qu

***Abstract* — A large amount of data is available on Twitter that can be used to manage different types of risks in financial institutions. This paper shows how machine learning algorithms can be applied to analyze large unstructured data and train a model to make a future prediction on tweets to categorize them by risk type and use sentiment analysis to understand the risk type. This model reads each tweet and categorizes them by risk using a specified dictionary and adds sentiment analysis to show the risk type seen in each tweet. Logistic regression used in this research helped to formulate the prediction model. Twitter data from 2019 was used to train and test a supervised machine learning algorithm and once the model started predicting tweets efficiently, it was used to predict twitter data from 2022 in our experimental research. Our experiment confirmed that Twitter data can be used to manage risk with the right type of modeling using machine learning techniques.**

***Keywords* — Artificial Intelligence; Machine Learning; Natural languages Processing; Text Analytics.**

I. Introduction

Data is a treasure that can help businesses to improve their decision-making and performance to stay competitive in this current economy. The daily amount of data we get from different sources including social media is huge, trying to categorize and analyze the data manually is next to impossible. In such situations, artificial intelligence can be a useful solution to solving such problems. Machine learning, which is a part of artificial intelligence, is especially effective for solving problems with a fundamental statistical component like regression, classification, and clustering, which is assumption-free and non-parametric [1]. Three types of machine learning are available for different scenarios: supervised, unsupervised, and reinforcement. A supervised machine learning algorithm works with a set of presented data where the input corresponds to the output to find the relationship between the input and output [2]. Moreover, supervised machine learning is mostly used in classification problems to predict values or regression problems to predict the continuous-valued output.

This paper presents how financial institutions can manipulate Twitter data to understand their competitors and their customers' experience which can help them to identify the types of risk they may be facing. Statista [3] shows as of 2021, Twitter had 8.85% of the world's overall social media user base; with 397 million users and most of the users are from the United States (73 million) followed by Japan (56 million) and India (22 million); top use reason for news source (48%), mostly by age group between 25-49 (59%). Since twitter data are large and unstructured, financial institutions can apply machine learning techniques using natural language processing (NLP) with a python programming language to categorize risk and find the sentiment of each risk category.

Some of the risk types that concern financial institutions are their compliance with regulations, competitiveness with other financial institutions, and their performance in relation to financials or customer experience. After the Dodd-Frank Act was imposed in the United States, it requires enhanced data capabilities in data sourcing, data processing, data retention, analytics and reporting, data management, and data governance and control. As explained by Krishna [4], risk management has evolved significantly after the Dodd-Frank Act became law by requiring larger datasets, increased data processing, data integration across organizations, and timeliness of data. Data collected from Twitter can be utilized to manage risk in financial institutions with the help of text analytics; competitive risk, compliance risk, and performance risk have shown a direct correlation with bank risk rate [5] This paper will focus on how to analyze and categorize Twitter data using supervised machine learning to predict future tweets and classify them by risk type.

II. Related Works

A. Natural Language Processing

In today's marketplace, data is a treasure that can help decision-making perform well and stay competitive. Social media data can be used to understand trending sentiments around a particular topic in real-time [6]. Financial institutions can identify different types of risks from social media using natural language processing (NLP) to monitor compliance risk, competitiveness, and performance [7]. This open-source data can help financial institutions to adjust their risks depending on what they witness in the real world. We can read individual tweets and understand what the person states on that tweet and try to understand how they felt about a particular experience, but to aggregate multiple tweets and analyze the sentiment or categorize it by risk type manually is very time-consuming. The Python programming language enables the performance of the processes of tokenization, stemming, and stop-word more quickly using preloaded libraries [8]. Furthermore, Python has also been considered

Submitted on March 06, 2022.
Published on March 31, 2022.
I. M. Haile, Colorado Technical University, USA.
(e-mail: imhaile27@gmail.com)
Y. Qu, Colorado Technical University, USA.
(corresponding e-mail: yqu@coloradotech.edu)

the best language for learning to program as it reflects the way people think in its code implementation.

There are different types of Python software packages to analyze sentiments in social media, one of these packages is Valence Aware Dictionary and Sentiment Reasoner (VADER). VADER helps to identify which tweet is positive, neutral, or negative [9]. VADER is a sentiment analysis that is validated by humans from a generalizable, valence-based, and human-curated sentiment lexicon.

B. Machine Learning

Financial institutions use data mining for marketing, fraud detection, risk management, and investment banking to discover knowledge using machine learning, information science, visualization, and statistics [10]. According to Kaddouri [11], what differentiates data mining techniques from the traditional statistical methods is that it is free from hypothesis by training and validating multiple approaches. In addition, data mining should include human expertise in the process to improve the value of data mining technology. As stated by Fry [12], artificial intelligence is a great prediction tool based on gathering data from different sources that are not possible by humans. Artificial Intelligence needs to fit the formats and workflows of the users of what they want to accomplish [13]. Depending on where AI is used, it is crucial to align it with the organization's goals and support the user's experience. There are three types of machine learning: supervised, unsupervised, and reinforcement. A supervised machine learning algorithm works with a set of presented data where the input corresponds to the output to find the relationship between them [2]. Moreover, supervised machine learning is mostly used in classification problems to predict values or regression problems to predict the continuous-valued output. Machine learning is a statistical algorithm used to solve problems that has an underlying statistical component like regression, classification, and clustering which is assumption-free and non-parametric [1]. Algorithms in machine learning can be used to train, validate, and test data in computers to create predictive models that can be refined until it becomes valid for real situations [14].

C. Risk Management

Risk management has evolved in recent years as regulators focus on the financial industry, which required large datasets from multiple sources to analyze risk [4]. Some of the risk types that interest financial institutions are their compliance with regulations, competitiveness with other financial institutions, and their performance in relation to financials or customer experience [7]. Krishna [4] contended that regulators are responsible for the transformational changes experienced in the banking sector regarding reporting hierarchies, organizational structure, and adopting technology required to monitor and report risk. Data collected from Twitter or other social media sources can be utilized to manage risk in financial institutions with the help of text analytics; competitive risk, compliance risk, and performance risk have shown a direct correlation with bank risk rate [5]. If Twitter data analytics is applied well in the financial sector, it can be a great resource for strategic decision-making and mitigating risk.

III. Problem Statement, Hypothesis Statements, And Research Questions

A. Problem Statement

Financial institutions use Twitter to share their performance and other activities on a daily basis, their customers also tweet about their experiences with their financial institutions. This data provides a good amount of information to understand customer experience and how the financial institutions are performing. However, Twitter data are unstructured and make it difficult to categorize a large amount of tweet data by risk type or sentiment. A different method should be used to quickly collect tweets and categorize them by their risk type and sentiments to learn from the data and make the necessary action by the financial institutions. Therefore, automation is important to group tweets by sentiments and by risk type to understand competition, compliance risk, and performance risk.

B. Hypothesis Statements

If financial institutions can utilize machine-learning algorithms to monitor social media like Twitter to find what is being discussed about their organization or their competitors, they can proactively mitigate their risks by applying the right strategies. Twitter users comment about their experiences with their financial institutions on a daily basis and financial institutions also tweet different topics related to their performance or accomplishments. Aggregating this information and analyzing it can provide helpful insight into each bank's performance, competitiveness, and meeting its compliance requirements. In order to improve in these risk areas (performance, competitiveness, and compliance) or other risk categories, financial institutions may be able to analyze Twitter data and understand which risk area they may need to improve on.

C. Research Questions

Based on the problem statement this research paper tries to answer the question: how can financial institutions use a machine-learning algorithm to categorize Twitter data based on risk types and sentiments at the same time to identify what type of risks they are facing?

IV. Methodology

The daily volume and velocity of twitter data are huge, trying to categorize and analyze the data manually to understand risk types or sentiment is next to impossible. In such situations, machine learning can be a useful technique for solving this problem. Machine learning is a statistical algorithm used to solve problems that has an underlying statistical component like regression, classification, and clustering which is assumption-free and non-parametric [1].

The logistic regression model will be used to help in prediction as this statistical model plays a significant role in predictions. Logistic regression can be applied to categorize multiple independent variables into either two or more mutually exclusive classes that contain exploring a linear mixture of regressors that confirms large variances in group means [15]. Logistic Regression in machine learning algorithms is favorable in classification to learn the association and categorical dependents versus independent

elements if dependent elements have binary values like 0 and one, true or false, positive, or negative, or yes or no [16]. The equation for logistic regression below provides the idea of how dependent elements and independent elements of the dataset:

$$i = Logistic\ Regression\ (p) = 1n\ (p/\ (1-p))$$

The research method selected for this study is experimental testing of quantitative data with the use of machine learning techniques that applies a logistic regression model. Algorithms in machine learning can be used to train, validate, and test data in computers to create predictive models that can be refined until it becomes valid for real situations [14]. The sample data collected from Twitter will be used to train and test the machine learning algorithms to make a prediction on new tweets collected in the future. Logistic regression can be applied to categorize multiple independent variables into either two or more mutually exclusive classes that contain exploring a linear mixture of regressors that confirms large variances in group means [15].

For training and testing this machine learning algorithm, 132,767 twitter data were used from 2019 that mention financial institutions in the United States such as Citi, Goldman Sachs, JP Morgan, U.S. Bank, and Wells Fargo. Using python programming these tweets were tokenized to separate the significant words to help the categorization and sentiment analysis. Once the tokenization is finalized, a dictionary with 1,582 words is added. The dictionary was created by researching for words that correspond to each risk category.

TABLE I: COUNT OF DICTIONARY WORDS BY RISK TYPE AND EXAMPLES

	Compliance	Performance	Competitive	Total
Word Counts	433	530	619	1,582
Examples	Lawsuit, governance, disclose, violation…	Profit, earning, downgrade, plan…	Better, introduce, modify, prize…	

The program looks for the tokenized words from Twitter and tries to match the word from Twitter to the dictionary when it finds a match, it categorizes that tweet as compliance, performance, or competitive depending on which column the word was identified in. In addition to categorizing the tweet by risk type, the VADER sentiment analyzing package was applied to identify the sentiment of each tweet if it is positive, negative, or neutral. VADER is an open-source lexicon and rule-based sentiment analysis tool designed to analyze sentiments expressed in social media. VADER is a sentiment analysis that is validated by humans from a generalizable, valence-based, and human-curated sentiment lexicon [9].

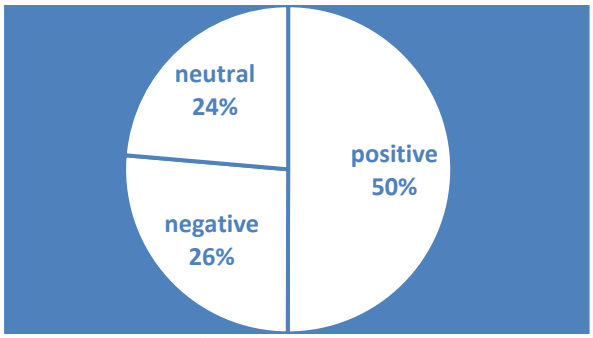

Fig. 1. Results of Sentiment Analysis on Training Data.

The sentiment analysis completed on our total data shown in figure 1 result, that 50% of the tweets are positive followed by 26% negative tweets, and the remaining 24% shows as neutral tweets. Analyzing the sentiment after categorizing the risk type helps if a tweet in that category needs an action plan by the financial institutions. For example, if we have a tweet that is categorized as compliance and positive, it may not need an action plan, but if we have a tweet that shows a negative compliance tweet, the bank mentioned on that tweet should have an action plan to mitigate that risk

TABLE II: FREQUENCY OF RISK CATEGORY

Category	Frequency
Competitive	44794
Performance	41584
Neutral	38727
Compliance	7662
Total	132767

Once each tweet is assigned with the risk type and sentiment type, the data is ready for training and testing. Out of the total data for this experiment, 70% was used to train the machine learning algorithm, and the remaining 30% was used for testing the trained algorithm. The accuracy result of the logistic regression returned 92.85% which is close to 100% and can be used for future prediction. Logistic regression is a good way to estimate the relationship between one or more independent variables and binary outcome variables and more formally it can be used to estimate the probability (or risk) of a given value [3]. This accuracy result received on our data allows us to proceed with using the trained model to predict future tweets and categories them with risk type and add a sentiment.

Now that we have our model trained and tested that shows a good accuracy result, the next step is to pull additional Twitter data to make predictions. For this experiment, 150 new data from 2022 was pulled that mentions the word “bank” to include any tweet that discusses financial institutions. Python program looks at the data on the new file and uses the model that we trained and tested to make the prediction. The resulting file on the new 2022 Twitter data now shows its prediction of each tweet’s risk type and sentiment using the model that was trained with 2019 data.

TABLE III: EXAMPLE OF TWEET RESULT BY CATEGORY AND SENTIMENT

Date	Tweets	Tokens	Category	Sentiment
7/6/19	A similar dynamic plays out during the financial...	['similar', 'dynam', 'play', 'financi', 'crisi...	performance	positive
12/23/19	JPMorgan Chase's expansion and the mega-merger...	['jpmorgan', "chase'", 'expans', 'mega-merg', ...	performance	positive
8/12/19	#GoldmanSachs said on Sunday that fears of the...	['goldmansach', 'said', 'sunday', 'fear', 'u'...	competitive	negative
1/15/19	@Ask_WellsFargo @Dillards If I want a Dillard'...	['want', "dillard'", 'credit', 'card', 'specif...	neutral	positive
2/20/19	their silly terms and conditions say they don...	['silli', 'term', 'condit', 'say', 'warn', 'an...	competitive	negative

V. Experiment Results

Using Twitter data from 2019 that mentions major financial institutions in the United States such as Citi, Goldman Sachs, JP Morgan, U.S. Bank, and Well Fargo a machine learning technique was developed that applies a logistic regression model to categorize tweets by risk type and include sentiment analysis. The program looks for the tokenized words and the dictionary if there is any match and when it finds one, it categorizes that tweet as compliance, performance, or competitive. In addition, the VADER sentiment analyzing package was applied to identify the sentiment of each categorized tweet if it is positive, negative, or neutral.

A logistic regression model was applied to train and test the data to find the accuracy result received on our data to allow us to proceed with using the trained model to predict future tweets and categories them with risk type and add a sentiment. The model shows an accuracy of 93.23% ability that it can predict the three risk category types and assign one risk type for each tweet. The adjunct sentiment helps in identifying if the tweet or the risk type is positive or negative to help users depending on what they need to focus on. Logistic regression is useful to estimate the relationship between one or more independent variables and binary outcome variables and more formally it can be used to estimate probability or risk on a given value [17].

TABLE IV: Prediction Result Examples on New Tweets

Date	Tweet	Category	Sentiment
1/1/2022	$300 Giving in 72 HOURS ðŸ‘¾ðŸŽ¾ RT & @FreaksNGuilds RT Recent Tweets And Like + Tag @TheShamdoo & @Banks	Competitive	Neutral
1/1/2022	US banks have been more proactive than other industries in encouraging employees back to the office, but those plans have come under renewed scrutiny due to the rapid spread of the Omicron coronavirus variant.	Performance	Negative
1/1/2022	Would you say that being the primary U.S. large bank regulator, and then leaving that role and accepting $7.2 million in just 2 years from those same banks in "speaking fees," is corrupt? https://t.co/Kzurw3FT8Q	Compliance	Positive
1/1/2022	@Arron_banks Brexit is costing us all so much Â£.	Performance	Negative
1/1/2022	banks owned a 30% share in Weibo. That's why they were able to censor pro-China anti-western contents for decades. However, things are changing... Alibaba is selling its shares. The good old days of the fifth column is over.	Performance	Positive

The new 150 twitter data pulled from 2022 mentions the word “bank” to include any tweet that talks about financial institutions. Python program looks at the data on the new file and uses the model that we trained and tested to make the prediction. The resulting file on the new 2022 Twitter data now shows its prediction of each tweet’s risk type and sentiment using the model that was trained with 2019 data.

VI. Conclusion

Twitter is one of the major social media platforms with large users around the world, as of 2021, Twitter had 8.85% of the world’s overall social media user base; with 397 million users and the majority of the users are from the United States (73 million) followed by Japan (56 million) and India (22 million); top use reason for news source (48%), mostly by age group between 25-49 (59%) [3]. Financial institutions can use this data to manage their risk or identify new opportunities, but twitter data are large and unstructured. Therefore, financial institutions can apply machine learning techniques using natural language processing (NLP) with a python programming language to categorize compliance risk, performance risk, and competitive risk with sentiment analysis of each risk category.

In this research, we have shown how supervised machine learning can be used to categorize tweets by risk type and sentiment using logistic regression to train and test the algorithm using 2019 Twitter data. The model shows an accuracy of 93.23% ability that it can predict the three risk category types and assign one risk type for each tweet. The adjunct sentiment helps in identifying if the tweet or the risk type is positive or negative to help users depending on what type of risk they need to focus on. This model can be utilized in other social media by applying a similar method of training and testing to find similar risk types or other risk factors. Additional data training and testing can be done using Twitter data to increase the accuracy of the prediction.

References

[1] Ratner B. Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data. *CRC Press*; 2017 Jul 12.

[2] Al-Gethami KM, Al-Akhras MT, Alawairdhi M. Empirical evaluation of noise influence on supervised machine learning algorithms using intrusion detection datasets. *Security and Communication Networks*, 2021 Jan 15.

[3] Statista. Countries with the most Twitter users 2021. Published by Statista Research Department, Jan 28, 2022. Statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries

[4] Krishna D. Big Data in risk management. *Journal of Risk Management in Financial Institutions* 2016 Jan 1;9(1):46-52.

[5] Haile I. & QU Y. Quantitatively Examining the Relationship between Social Media Messages and the Risk Management at Financial Institutions. The *17th International Conference on Data Science* (ICDATA'21: July 26-29, 2021, USA).

[6] Park PH. Big data war: how to survive global big data competition. Business Expert Press; 2016 Aug 26.

[7] Haile IM. Data Analytics in Financial Institutions: How Text Analytics Can Help in Risk Management (Doctoral dissertation, Colorado Technical University).

[8] Quillo-Espino J, Romero-González RM, Paulin-Martinez FJ. Text mining preprocessing in Times of Python vs MVCS. International *Journal of Computer Science and Software Engineering*, 2019 Nov 1;8(11):266-75.

[9] Ribeiro FN, Araújo M, Gonçalves P, Gonçalves MA, Benevenuto F. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*,. 2016 Dec 1;5(1):1-29.

[10] Pulakkazhy S, Balan RS. Data mining in banking and its applications-a review.

[11] Kaddouri A. The role of human expertise in enhancing data mining. Capella University; 2011.

[12] Fry, A. The role of AI in decision making. AI and Big Data Expo, World Series. AI & Big Data Expo. February 2018. Retrieved from: Ai-expo.net/role-ai-decision-making.
[13] Hryniewicz, R. Three things CEOs should know about the use of artificial intelligence in decision-making. Cloudera, July 2018 Retrieved from: hortonworks.com/blog/three-things-ceos-should-know-about-the-use-of-artificial-intelligence-in-decision-making.
[14] Wanganga G, Qu Y. An Auto Optimized Payment Service Requests Scheduling Algorithm via Data Analytics through Machine Learning. In 2020 *International Conference on Computational Science and Computational Intelligence* (CSCI) 2020 Dec 16 (pp. 1498-1502). IEEE.
[15] Ali SS, Mubeen M, Lal I, Hussain A. Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX). *Asian Journal of Empirical Research*, 2018 Jul 6;8(7):247-58.
[16] Muhammad LJ, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN computer science*, 2021 Feb;2(1):1-3.
[17] Schober P, Vetter TR. Logistic regression in medical research. *Anesthesia and analgesia*, 2021 Feb;132(2):365.

Issayas M. Haile received his Business degree from National American University, Sioux Falls, SD, USA and earned his MBA and doctoral degree in Computer Science from Colorado Technical University, Colorado Springs, CO, USA.

In his professional career, he has worked at multiple financial institutions on different levels in operations, management, and consulting in areas like credit cards, mortgages, and other consumer products. He is currently Analytics Consultant for risk and controls at Wells Fargo Bank, NA. Previously published scholarly article with Springer Nature Group.

His current and future research interests are in data science, artificial intelligence, and visualization that can be applied in financial institutions or other industries to improve operations and risk management processes in an automated structure.

Yanzhen Qu received his B.Eng. degree in Electronic Engineering from Anhui University, China; M.Eng. degree in7 Electrical Engineering from the Chinese Science Academy, China; and Ph.D. degree in Computer Science from Concordia University, Canada.

Over his industrial professional career, he has served at the various executive level management positions responsible for Product R&D and the IT Operation at several multinational corporations. He has led his multinational engineering teams successfully developed several the world first very large real-time commercial systems and technologies. He is currently the Dean and a Professor of Computer Science, Engineering and Technology with Colorado Technical University, Colorado Springs, CO, USA. He is also the dissertation supervisor of many Computer Science doctoral students. He and his doctoral students have published several dozen scholarly articles, some of them received the best paper award at several IEEE international conferences.

His current research interests include data science, cybersecurity and privacy, machine learning, e-learning technologies, software engineering, cloud computing, and affective computing. He has served as either the general chair, or the program chair, or a keynote speaker at many IEEE, ACM, ASIS, and IFIP international conferences or workshops. He is also an editorial board member of several professional peer-reviewed Computer Science or Information Technology journals.