Icon

Description automatically generated UNIVERSITY OF WESTMINSTER

MSc Data Science and Analytics

Module: Web and Social Media Analytics

Module Code: 7BUIS025W

Course Work

Assigned By: Philip Worrall

Prepared By: Mohamed Shiban Lal

Student ID: 19281971

Submission Date: 2nd May 2023

Table of Content

Company Background ……………….......3

Data Collection…………….…...4

Exploratory Analysis ………………….……….7

Topic Model Analysis ……………………10

Sentiment Analysis…………………………14

Conclusion………………………17

Reference …………………….20

Appendix……………21

**Company Background**

Zero-emission electric cars (EVs) and related technologies are the focus of Nikola Motors, a company that is publicly traded. Trevor Milton founded the company in 2015 with a vision of modernising the transportation industry and reducing carbon emissions. With its global operations and USA headquarters, Nikola Motors aims to offer innovative and sustainable modes of transportation.

The company's main areas of developing focus are the creation of battery- and hydrogen-electric trucks and utility vehicles. The Nikola One, a long-haul hydrogen-electric semi-truck, and the Nikola Tre, an electric truck made for the European market, are two of Nikola's products. In order to attain long-range capabilities with zero emissions, these vehicles are powered by innovative electric powertrain systems that combine batteries and hydrogen fuel cells.

Along with selling vehicles, Nikola Motors also works on the development of fuel cell technology and a charging network to support its fleet of electric trucks. The company has joined forces with titans of the sector to improve hydrogen fuelling technologies and build out an extensive system of hydrogen stations.

*Relative Facts*

Strategic connections: To benefit from their expertise in fields like fuel cell technology, vehicle manufacturing, and hydrogen infrastructure development, Nikola Motors has established partnerships with companies like Bosch, CNH Industrial, and Nel Hydrogen.

Market Expansion: The company has extended its activities outside of the US to penetrate the European and international markets with its innovative EV solutions. To have a significant footprint in these areas, Nikola Motors has started partnerships and collaborations.

Financial outcomes: To support its goals, Nikola Motors has attracted significant funding as well as investment. In 2020, the business went public through a merger with a special purpose acquisition company (SPAC), garnering significant funding for its expansion and product development initiatives.

Nikola Motors is dedicated to improving sustainability and lowering carbon emissions in the transportation industry. The company wants to offer zero-emission vehicles that help create a brighter future by concentrating on battery-electric solutions and hydrogen fuel cell technology.

Despite some problems and difficulties, Nikola Motors has become well-known and prominent in the automobile industry. Investors, clients, and industry experts are all taking an interest in the company's innovative approach to electric vehicle technology.

These facts show Nikola Motors' commitment to revolutionise the transportation industry by providing innovative electric vehicles and promoting environmentally friendly options. Due to its focus on battery-electric and hydrogen fuel cell technologies, the company is positioned to play a significant role in the worldwide movement towards zero-emission transportation.

**Data Collection**

Keywords or search terms used for data collection are.

* Nikola Motors
* Trevor Milton resignation
* Nikola Motors Hindenburg Report
* Nikola Motors Scandal

*Approach Taken*

The method of data collecting that was adopted focused on significant issues and events associated with Nikola Motors, such as Trevor Milton's leaving, the Hindenburg Report, and the Nikola Motors controversy. This strategy was chosen to gather relevant and latest discussions and opinions on this company. With the help of targeted keywords and search terms.  We tried to gather social media posts and comments that expressly highlighted these subjects. This strategy enables us to focus on significant occurrences and conversations that have influenced the general public's opinion of and attitude toward Nikola Motors.

Additionally, by combining data from Reddit and YouTube, we can compile a wide range of user-generated content about Nikola Motors, including discussions, reviews, and video content. While YouTube offers video content and related comments, Reddit offers a platform for users to participate in discussions and share their thoughts in text-based posts and comments.

*Step followed to collect social media data*

* Chose Reddit and YouTube relevant Social Media platforms to use.
* Using APIs and web scraping tool (Google Colab) Python to extract user comments from Reddit and YouTube content linked to Nikola Motors.
* Focus on Reddit and YouTube as the main information sources while compiling Nikola Motors-related social media content. Reddit offers text-based discussions and points of view (Comments), while YouTube has videos and comments on them.
* Created search queries and filters based on the chosen keywords, such as "Trevor Milton," "Nikola Motors," and other relevant terms. To make sure that the data fits with the analysis's specific focus, these filters and queries help to focus the data.
* Based on the above keyword, searched for the relevant post on Reddit and YouTube and created queries to gather data such as (comments) which was executed in the relative Reddit and YouTube collecting data code given in [Appendix - A](#_Appendix).
* After generating the data, the data were combined and imported in a codec function that read the line greater than 10 characters, applied UTF-8 encoding which handles text containing non-ASCII characters, and stored in a CSV file structured format that is suitable for analysis. This makes manipulating, exploring, and further processing the data simpler.

To make sure that the data collected reflects recent discussions and the current sentiments about Nikola Motors, a 7-day time filter was used. This period finds a compromise between collecting current data and providing enough volume for analysis.

*Data Pre-processing Steps and Justification*

* *common\_case():*The text is converted to lowercase using this function. By lowercasing the text, you may ensure consistency and get rid of any discrepancies caused by using multiple capitalizations. It improves text analysis and helps in avoiding repetition.
* *no\_multi\_punctuation():* This function removes multiple instances of punctuation, including hashtags, colons, question marks, exclamation points, and at symbols. Repeated punctuation marks should be removed to make the text cleaner and reduce noise in future analyses.
* Using the function *without\_leading\_trailing\_whitespace(),* any leading or trailing whitespace is taken out of the text. Whitespace in the leading and trailing directions may cause additional variation which decreases the accuracy of studies. This whitespace must be eliminated to ensure consistency in the text data.
* The "@" sign, which normally denotes a retweet or reference of another person, is removed away by the function *no\_retweets().* By deleting retweets, the analysis focuses on user-generated discussions and unique content rather than retweeted stuff.
* *no\_http\_links():* removes words that begin with "http," which are probably URLs or links to websites. By removing links, it is possible to analyse text content independently of external online references.
* The input text is cleaned using several pre-processing techniques by the *preprocessing\_pipeline()* function. The text is changed to lowercase, leading and trailing whitespace is removed, consecutive punctuation is removed, retweets are filtered out, and website URLs are removed.

Graphical user interface, text, application, email

Description automatically generated

* The preprocessing\_pipeline2() function is changed in the code snippet so that it now works with rows of DataFrames rather than individual text inputs. This enables the preprocessing procedures to be applied on the DataFrame's "text" column.

Graphical user interface, text, application

Description automatically generated

* Following the application of the pre-processing steps, the code measures the length of each cleaned comment and keeps only comments that are 10 characters or longer. This filtering assists in removing extremely brief or meaningless comments that might not offer insightful discussion.

Graphical user interface, text, application, email

Description automatically generated

* Next, uses the langdetect library to determine each comment's language next. The Data Frame is further filtered to only keep the comments recognised as English ("en") after applying language detection to the cleaned comments using the *language\_code()* function.

Graphical user interface, text, application, email

Description automatically generated

These methods of pre-processing serve to clean the text data, get rid of extraneous information and irrelevant content, and guarantee consistency for further analysis. Each step is justified by improving the dataset's consistency and quality, making the results easier to understand, and concentrating the analysis on English-language comments related to Nikola Motors.

**Exploratory Analysis**

*Approach Taken*

With the help of the code snippet in [Appendix - B](#_Appendix), exploratory analysis is done on the social media information collected from the "combined\_comments.csv" file. To obtain insights into the data, it begins by reading the data into a Pandas Data Frame and then performing various operations on the Data Frame.

The comments by author feature display a bar graph of the top 15 authors' comment posting volume. The most active participants in the conversations around Nikola Motors can be found using this analysis.



To maintain data integrity, all rows with missing values were removed using the dropna() function.



Used the Counter class from the collections module to determine the frequency of each word in the cleaned-up comment text. The top 10 words and their frequencies are gathered into a Data Frame. The most commonly occurring words in the comments have been identified by this analysis. [Appendix - B](#_Appendix)

A picture containing graphical user interface

Description automatically generated

Using the WordCloud library to visualise word clouds. The word cloud displays the 200 words from the comment text that are used the most frequently, with the size of each word representing its frequency. The terms that are prominent in the comments are represented in this visualisation in a visually pleasing way.

Graphical user interface, text, application

Description automatically generated

*Appropriate Visualisation*

A bar graph shows how many comments the top authors have made. The writers are shown on the x-axis, while the quantity of comments is shown on the y-axis. This visualisation can be used to determine the discussion people who are the most active.

Chart, histogram

Description automatically generated

To visualise the most frequently used words in the comments, a word cloud is created. Each word's magnitude in the cloud corresponds to how frequently it appears in the comment text. The important themes or keywords in the comments are clearly displayed in this visualisation.

Text

Description automatically generated

*Critical Evaluation*

The techniques used in the exploratory analysis code, such as word cloud visualisation, bar charts, and word frequency analysis, are commonly used in exploratory data analysis of text data. These methods offer insightful analyses of the data and help with finding of similarities, trends, and unique characteristics.

Understanding the most active participants in the conversations is made possible by the bar chart of the top authors. It's crucial to remember that the number of comments may not necessarily reflect the value or relevance of the contributions. Additional analysis, like sentiment analysis, can offer more in-depth perceptions of the sentiments expressed by various authors.

The most frequently used words in the comments are highlighted by word frequency analysis and word cloud visualisation. It is crucial to keep in mind that stopwords are eliminated during the pre-processing stage, which could affect the inclusion of some common words. Furthermore, because word frequency is used in the visualisation, the context or semantic meaning may not be captured.

*Initial Result & Discussion on Finding*

According to the initial results of the exploratory analysis, as shown in the bar chart, it can be seen that some authors are very active in the discussions regarding Nikola Motors. This suggests the existence of important decision-makers or loyal followers.

Insights on the most common topics or keywords mentioned in the comments are provided by the word frequency analysis and word cloud visualisation. These results can be used to pinpoint the key topics of discussion and areas of participant interest.

To fully understand how the public feels about Nikola Motors, it is crucial to go deeper into the feeling that was expressed in the comments. To determine the negative, neutral, and positive attitudes represented in the comments, sentiment analysis might be used. This can give a more complex picture of how people feel about Nikola Motors and its products normally.

It would also be advantageous to thoroughly examine the context and content of the comments. For instance, recognising specific subjects or issues raised, spotting patterns or trends over time, and investigating relationships between sentiment and certain Nikola Motors-related news or events.

Overall, the initial results from the exploratory analysis provide an outline for more investigations. The public's perception of Nikola Motors can be better understood by using topic model analysis and sentiment analysis.

**Topic Model Analysis**

*The methodology used to conduct analysis*

* The code uses Latent Dirichlet Allocation (LDA) topic modelling using the gensim package.
* The pre-processing of the comments involves storing them into individual words to create a dictionary.
* The encoded comments are transformed into a bag-of-words format to create the corpus.
* The corpus with a certain number of topics is used to train the LDA model.
* The subjects are displayed and printed.

Graphical user interface, text

Description automatically generated

*The result, Visualisation, and distribution of comments*

The LDA model's recognised patterns are printed by the code above. A collection of terms with their accompanying weights act as a model for each topic. Here is an example of the output.

Text, letter

Description automatically generated

Normalised Score Histogram: This histogram displays the normalised scores derived for each comment's frequency distribution. It helps in the understanding of the comments' general sentiment.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Sentiment Score Histogram: This histogram displays the frequency distribution of each comment's sentiment score. It provides a broad picture of the remarks' emotional range.

Chart, histogram

Description automatically generated

Distribution comments over topic Histogram: The number of comments assigned to each topic is displayed in this histogram. It is useful for understanding how comments are distributed among the different subject matter.

*Evaluation*

Goodness of Fit and Perplexity:

The code uses the CoherenceModel class from gensim to determine the coherence score. The coherence rating evaluates how easily topics can be understood. A better topic definition is indicated by a higher score.

Text

Description automatically generated

The LDA model's Perplexity is also calculated. It evaluates how accurately the corpus is predicted by the model. Better performance is indicated by lower values.

Graphical user interface, text, application

Description automatically generated with medium confidence

*Interpretation of result and discussion of findings*

* The LDA model's topic predictions shed light on the primary topics covered in the comments.
* The sentiment analysis helped to identify the overall attitude—whether positive, negative, or neutral—expressed in the comments.
* The distribution of comments across participants sheds light on the relative importance of various topics and helps in the discovery of the data's strongest topics.
* The evaluation metrics, such as perplexity and coherence score, gave insights into the efficacy and effectiveness of the topic model approach.

*Approach of Critical Evaluation*

* Topic modelling and sentiment analysis on the comment data is done using the methodology used in this analysis.
* While sentiment analysis helps understand the expressed sentiment, the LDA model helps in identifying the underlying topics in the comments.
* The evaluation metrics offered some insight into the model's effectiveness and quality.
* Topic modelling and sentiment analysis methodologies have some limitations, nevertheless, which should be considered. The quality of the data and the model's underlying assumptions determined how accurate the results are. The results must be independently verified and carefully interpreted.

Overall, the analysis offers insightful information on each topic covered in the comments, the approach used, and the distribution of comments among various topics. Based on the analysed comments, it helps in understanding how Nikola Motors is seen by the general population.

Finding the main concepts or topics that dominate the comments is made easier by the topic modelling analysis using LDA. It is possible to comprehend the topics being discussed more thoroughly by looking at the word distributions within each topic.

The sentiment analysis helps to determine the underlying sentiment behind the comments. It is possible to tell whether a comment is positive, negative, or neutral by computing sentiment scores for each one.

The distribution of comments over topics helpful in determining what topics are most frequently found within the data. Finding the topics with the most participation or discussion can be done by visualising the number of comments each topic has received.

The evaluation metrics, including perplexity and coherence score, offer an evaluation of the effectiveness and value of the topic's modelling approach.

As a result of the study of the comments, subject modelling and sentiment analysis offer insightful information about how the public views Nikola Motors. The analysis helps in locating significant topics, understanding the language used, and examining how remarks are distributed among different topics. These results may inspire more research, decision-making, and communication strategies concerning Nikola Motors.

**Sentiment Analysis**

*Reflection on the Role of sentiment analysis*

Sentiment analysis is crucial for social media modelling because it enables users to communicate their sentiments or opinions in posts, comments, or reviews. It gives us insights into how people feel about trends, public opinion, and customer satisfaction and allows us to quantify the polarity of the sentiment (positive, negative, or neutral).

*Methodology Used*

The NLTK library's VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon was used in the sentiment analysis process. To determine the sentiment of a given text, this lexicon-based approach makes use of a preset collection of terms and the sentiment scores that correspond to them. Based on the presence of particular words, intensifiers, negations, and other linguistic features, the sentiment scores are determined.

The DataFrame's "cleaned\_body" column, which had the preprocessed comments, was used for the sentiment analysis. The sentiment scores for each comment were determined using the SentimentIntensityAnalyzer from NLTK's VADER module. The sentiment scores were then stored in the DataFrame's new "sentiment\_score" column.

Text

Description automatically generated

*The result, and Visualisation*

A histogram of the sentiment scores was generated to show the frequency distribution of the sentiments and to express the results. Additionally, a histogram of the normalised scores was used to normalise the sentiment scores in order to better understand the distribution.

Text

Description automatically generated

*Goodness of fit and other diagnostic indicators*

The provided code does not clearly evaluate the quality of fit or any other diagnostic indicators. However, metrics like accuracy, precision, recall, and F1-score can be calculated by comparing the sentiment analysis results with manually annotated ground truth data in order to evaluate the performance of sentiment analysis. These metrics offer an evaluation of the sentiment analysis model's accuracy and reliability.

*Interpretation of result and discussion of findings*

Analysing the sentiment score distribution and understanding the general sentiment trends are necessary for interpreting the findings. It helps in determining the most common sentiment—whether positive, negative, or neutral—and can shed light on how the public feels about Nikola Motors. For instance, a high frequency of positive scores indicates a favourable perception, whereas most positive ratings indicate a positive perception.

*Method and technique used for Critical Evaluation*

It is crucial to recognise the limits of lexicon-based approaches like VADER while evaluating methods and procedures. These methods heavily rely on the availability and accuracy of the sentiment lexicon, which might not fully capture details and sentiments that are unusual to an instance. Additionally, lexicon-based approaches might have trouble classifying texts that contain sarcasm, irony, or ambiguity.

The limits of sentiment analysis in social media models must also be taken into consideration. Sentiment analysis can be difficult with social media data since it can be chaotic, unstructured, and context-dependent. Advanced methods like contextual analysis and machine learning models are needed for understanding the context, sarcasm, and cultural references.

In conclusion, sentiment analysis is crucial to social media modelling since it sheds light on the sentiment and opinions of people in general. NLTK's VADER module was used together with a lexicon-based methodology in this case. Sentiment scores and their distributions were visualised to show the results. However, diagnostic indicators and the evaluation of goodness of fit were not provided. Given the nature of social media data and the potential drawbacks of lexicon-based approaches, it is crucial to thoroughly evaluate the capabilities and effectiveness of sentiment analysis tools.

**Conclusion**

*Key Result*

Exploratory Data Analysis: The analysis determined the top comment-contributing authors and gave a visualisation of their comment-frequency distribution.

Pre-processing and Word Analysis: Stopwords, URLs, and emoticons were removed from the comments as part of the pre-processing process. The most common terms used in the comments were compiled into a word cloud, which gave an overview of the most frequently used words.

Latent Dirichlet Allocation (LDA) was used in topic modelling to identify the basic concepts in the remarks. The most popular topic and the keywords they were related to were retrieved, giving information about the major topics covered.

Sentiment Analysis: The sentiment scores for each comment were calculated using the VADER lexicon. Histograms were used to visualise the sentiment scores, showing how the sentiments were distributed and highlighting the polarity of the overall sentiment.

*Critical Evaluation*

The adopted approach has a number of advantages. The dataset was initially understood, and important contributors were identified using exploratory data analysis approaches. Stopwords, URLs, and emojis were successfully removed from the text data during the pre-processing processes. Understanding the prevalent trends and sentiment represented in the comments was made possible by-word analysis, topic modelling, and sentiment analysis.

There are some restrictions to consider, though. By considering additional text cleaning techniques, such as deleting special characters or handling misspelt words, the pre-processing stages could be improved still more. The coherence and interpretability of the identified topics could be measured using coherence scores to evaluate the topic modelling results. The accuracy and dependability of the sentiment analysis could also be assessed by contrasting it with manually annotated ground truth data.

Additionally, the VADER lexicon, which was used in the sentiment analysis method, might not have captured all nuances and context-specific sentiments. It's crucial to consider the drawbacks of lexicon-based methods for addressing irony, sarcasm, and ambiguous language.

In conclusion, the analysis was successful in revealing public opinion, mood, and major concerns surrounding Nikola Motors. The findings help to clarify the sentiments stated in the comments and the recurrent topics concerning the company. However, there remains room for development in terms of pre-processing methods, topic modelling evaluation, and more sentiment analysis approach validation.

*Intuitive explanation of the result*

Top Authors: By identifying the top authors who made the most comments, it determines the level of interest in and influences over the Nikola Motors scandal. These authors could be journalists, professionals in the field, or those who have been directly impacted by the incident. Their remarks can offer insightful information about how the public is perceiving and responding to current events.

The primary issues and discussions surrounding the Nikola Motors scandal are identified using the topic modelling analysis. These subjects might include dishonest acts, false advertising, corporate governance, or the effect on stakeholders and investors. The company and stakeholders can better understand these issues by being aware of the specific elements of the situation that the public finds disturbing.

Sentiment Analysis: The sentiment analysis gave a general summary of the polarity of the sentiments expressed in the comments on the Nikola Motors scandal. Due to the discussion and trust loss caused by the incident, negative emotions predominate. The sentiment analysis helps to identify the extent of public dissatisfaction or disappointment, which already had big impact on the company's standing and prospects in the future.

It makes intuitive sense that the analysis' findings in the context of the Nikola Motors fraud show a significant amount of criticism and negative sentiment. Undoubtedly, the scandal's charges of fraud and dishonest business practices have seriously damaged trust and damaged the company's reputation.

*Identification of potential next steps and avenues for future work*

Crisis management and reputation repair should be the company's main priority in order to deal with the incident, win back public confidence, and repair its reputation. This could involve open communication, accepting accountability for mistakes, taking corrective action, and exhibiting a dedication to moral behaviour.

Nikola Motors should interact with important stakeholders, such as investors, consumers, and the public, in order to hear their concerns, resolve issues, and create trust. This can be accomplished by having an open discourse, being honest, and showing concrete steps being taken to make things right.

Governance and Compliance Improvement: The crisis makes clear how important it is to have strong company governance procedures. To avoid such problems in the future, Nikola Motors must evaluate and strengthen its internal control systems, ethics policies, and governance structures.

Consider using impartial third parties or external auditors to verify the company's claims regarding its technology and financial procedures. This may help in regaining credibility and giving stakeholders reassurance.

Establish a strategy for ongoing social media and online platform monitoring in order to measure sentiment, identify new issues, and proactively resolve concerns. As a result, Nikola Motors will be able to monitor public opinion and emotions and take swift action.

Conduct an internal review and analysis of the scandal to determine its underlying causes, any lessons learned from them, and possible areas for improvement. This introspection will help the company create a culture of honesty and integrity and help it avoid similar problems in the future.

Accordingly, it is crucial for the company to concentrate on crisis management, reputation repair, stakeholder engagement, and governance improvement in light of the Nikola Motors scandal. Nikola Motors can work to restore its reputation and go forward with a fresh commitment to ethical conduct and transparency by addressing the issues brought up, re-establishing confidence, and setting corrective measures in place.

**Reference**

*Aggarwal, C. C. & Zhai, C. (2012) Mining text data. New York, N.Y: Springer.*

*Home* (2023) *Nikola Motor*. Available at: https://www.nikolamotor.com.

*John Rosevear (2022) Nikola founder Trevor Milton stands trial on fraud charges, capping a stunning rise and fall, CNBC. CNBC. Available at: https://www.cnbc.com/2022/09/13/nikola-founder-trevor-milton-stands-trial-on-fraud-charges.html.*

*Klender, J. (2020) Nikola motor labeled "an intricate fraud" by Hindenburg Research, TESLARATI. Available at: https://www.teslarati.com/nikola-motor-hindenburg-research/.*

*Press release (2021) SEC Emblem. Available at: https://www.sec.gov/news/press-release/2021-267.*

*Press, T.A. (2022) The founder of a zero-emission truck venture is found guilty of fraud, NPR. NPR. Available at: https://www.npr.org/2022/10/14/1129248846/nikola-founder-electric-trucks-guilty-fraud.*

*Russell, M. A. (2011) Mining the social web. First edition. Beijing ;: O’Reilly.*

*Worrall, P. (2023). Week 8,9,10,11,12: [lecture/tutorial notes]. Web and Social Media Analytics. Available from https://learning.westminster.ac.uk/ultra/courses/\_91254\_1/outline.*

# Appendix

Appendix - A

*YouTube Data Collection*

Text

Description automatically generated

*Reddit Data Collection*

Text

Description automatically generated

*Combining and Character handling*

Graphical user interface, text, application, email

Description automatically generated

Appendix – B

Graphical user interface, text

Description automatically generated with medium confidence

*Frequency Analysis of Word*

Text

Description automatically generated

*Removing Stopword*

Graphical user interface, text, application

Description automatically generated

*Normalised score*

Graphical user interface, text, application

Description automatically generated