Twitter sentiment analysis and named entity recognition.

Authored by: Anirban Das, Dominik Ktowera, Antoni Piotrowski
Date: 27th January, 2024.

Contents

List of figures

Introduction

With the introduction of the computer and the internet era at the end of 20th , the 21st century saw the rise of the online contents and data. As the internet is one of the main pillars of the modern society, it made enormous quantity of the machine readable documents get available online, thus becoming to be the most enriched database of knowledge if extracted properly and legally. From this data, various deep insights can be obtained which is generally not possible to depict relying on the humane analytical skills. This led to the invention and the adoption of the process of *Webs Scrapping* which aids in excavating and extracting textual data from online resources. This data is further analysed with the method of *Text mining*. The tech giant IBM[1] describes text mining as, "the process of analysing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends without requiring that you know the precise words or terms that authors have used to express those concepts." These key concepts, hidden themes, and relationships are obtained using various text mining techniques like Sentiment Analysis, Topic Modelling, Named Identity Recognition (NER), etc,.

In this paper the techniques of Sentiment Analysis is used to analyse the overall relation of tweeter user's sentiment on the Being Company for a given time period and the technique of the Named Entity Recognition is employed to check the most common entities related to these tweets for the Boeing company. The main goal is to check the sentiment of the investors and common public on social media platform and simultaneously excavate the other entities which are generally coming into conversation on the social media landscape. Concatenating the sentiment of the common public and the related entities together, there can be a possibility of extracting a comprehensive understanding of the sentiment trends of the investors and related key entities which can further be processed to use for machine learning prediction, empowering businesses, and assisting decision-makers gain insights without having to navigate the vast landscape of social media.

Literature Review

From the early stages of the introduction of the concept of text mining, sentiment analysis has always been a key aspect. Early 1990's research work slowly started with the sentiment analysis on classifying the semantic orientation of individual words or phrases, using linguistic heuristics or a pre-selected set of seed words (Turney and Littman, 2002)[2]. Then on 2002[3], Turney proceeded to classify document's sentiment into 'excellent' or 'poor' using the unsupervised techniques. This research was followed up by Pang et al. (2002)[4], who used the movie reviews data to classify the sentiment of the reviews, avoiding topic-based classification and used the supervised machine learning for the research work. Since then many different ways of sentiment analysis techniques evolved from Bag-Of-Words model (first coined by Zellig Harris, 1954), Valence Aware Dictionary and Sentiment Reasoner (VADER) scoring techniques by Hutto and Gilbert (2014)[5], to the advancement in the field of neural network (RNNs: LSTM) and development of the most recent Transformers, especially models like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and their variants, have demonstrated state-of-the-art performance on a wide range of natural language processing (NLP) tasks, including sentiment analysis. Similarly, the technique of Named Entity Recognition (NER) was used in mid-1990s by Grishman and Sundheim (1996)[6] and since then various approaches like Rules-based systems, Dictionary based systems, Supervised machine learning based systems evolved. Recently bidirectional architecture learning models like Bidirectional Encoder Representations from Transformers (BERT) in the field of Neural Networks has been very reliable in terms of accuracy in classification.

Methodology

This experiment is conducted using the twitter dataset obtained from the Github repository of the stocknet-dataset, added and updated by Yumo Xu. In this repository many different stock related tweets can be found, thus one of the main goal of this work is reproducibility of the codes. Though this paper is solely focussed on Boeing company ($BA) stock

related tweets, any other tweet dataset within this repository will be compatible with the code used for this analysis.

Data Preparation

The data preparation for this project was done slowly after running the pre-trained `SentimentIntensityAnalyzer` class from the Natural Language Toolkit `nlkt.sentiment` and analysing the scores from time to time, and enhancing the preparation.

In a step by step process, all the columns are evaluated and checked, finally dropping every columns except for the date ("created_at") column and the actual tweets ("text") column, containing 1104 entries. Then taking the textual data, urls, special characters, single characters with space, unnecessary spaces, unnecessary punctuations, along with the tickers ($BA in this case) is removed. The data is further tokenized and stop words are eliminated, finally preparing the data for the analysis.

Sentiment Analysis

The sentiment analysis was conducted on the prepared data using the Valence Aware Dictionary and Sentiment Reasoner (VADER) scoring technique. The pre-trained `SentimentIntensityAnalyzer` class from the Natural Language Toolkit `nlkt.sentiment` was used for this analysis. The scoring technique is used to evaluate the tweets and assign the corresponding sores to the adjacent column ("compound_sentiment_score").

Three word clouds are prepared for analysis from the positive (score $>$ 00.5), negative (score $<$ 0.05) and neutral (-0.05 $>$ score $<$ 0.05) scored tweets to check the most frequent words leading to these sentiment. Another important aspect i.e. Temporal Sentiment Analysis was taken into account and is plotted to check the valuable insights on how the sentiment changes over different time periods and that can be used to observe trends, patterns, or specific events that influence sentiment. Finally the tweets with highest and lowest scores are checked.

Named Entity Recognition (NER)

Named Entity Recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

`Spacy`'s NER model, specifically the `en_core_web_sm` is a small English model trained on written web text (blogs, news, comments), utilizes a statistical machine learning system. The system usually consists of a series of models that predict which parts of a sentence are names of things, such as companies or places, and what type of name they are (e.g., a person name vs. a company name). `Spacy`'s pipeline processes text in several steps, one of them being the NER which uses a combination of convolutional neural networks (CNNs) for tagging entities. The CNNs are used to extract features from the text.

Results

Sentiment Analysis

After the tweets were run using the `SentimentIntensityAnalyzer` class, the VADER scores are obtained. The distribution of VADER score is plotted below:

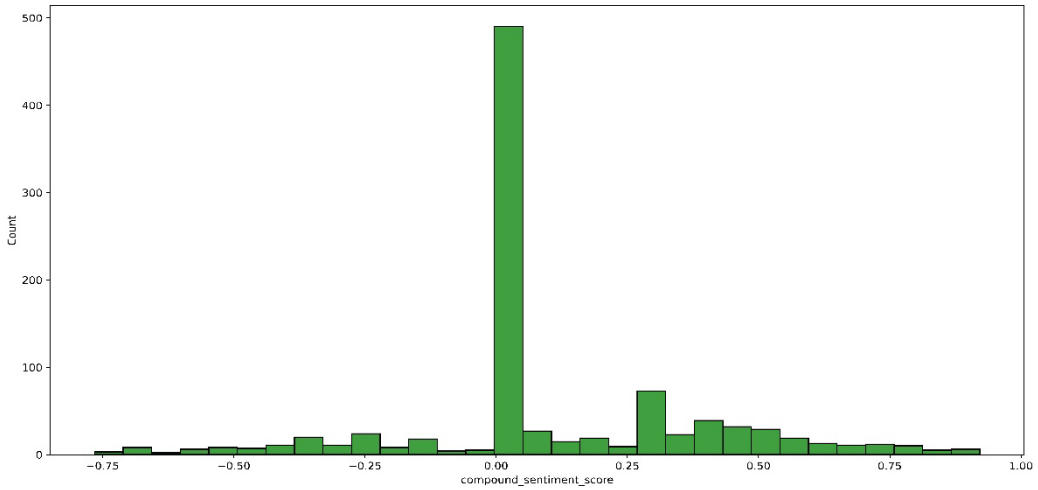

fig 1. Distribution of the sentiment scores <-1,1>

The distribution is normal with high numbers of the zeroes which is the ‘Neutral’ sentiment and the rest of the scores relates to ‘Positive’ and ‘Negative’ sentiment on both sides of the neutral sentiment.

fig 2. Positive tweets Word Cloud

The fig 2 refers to the word cloud for the positive tweets. There are very significant words like “earnings”, “trade”, “new”, “profit”, “win”, “gain”, “order”, “earnings”, etc., depicting the reason for the tweets being marked as Positive sentiment.

fig 3 Negative tweets Word Cloud

The negative word cloud depicted in fig 3 showing the frequency of the words that made the tweet classified with negative VADER score. The words like "low", "beat", "Malaysia" (referring to the Malaysian airlines "crash"), "problem", "drop", "cut", etc., made the tweets about the Boeing company and the $BA stock classified as Negative sentiment.

The above mentioned word clouds can be further enhanced by removing the topics related to the "boeing" "company" like "stock", other tickers like "ibm", "aapl", "spy", etc,. These are frequency of the words that come in both positive and negative tweets which can wither be ignored or further enhanced to only depict the words related to the sentiment.

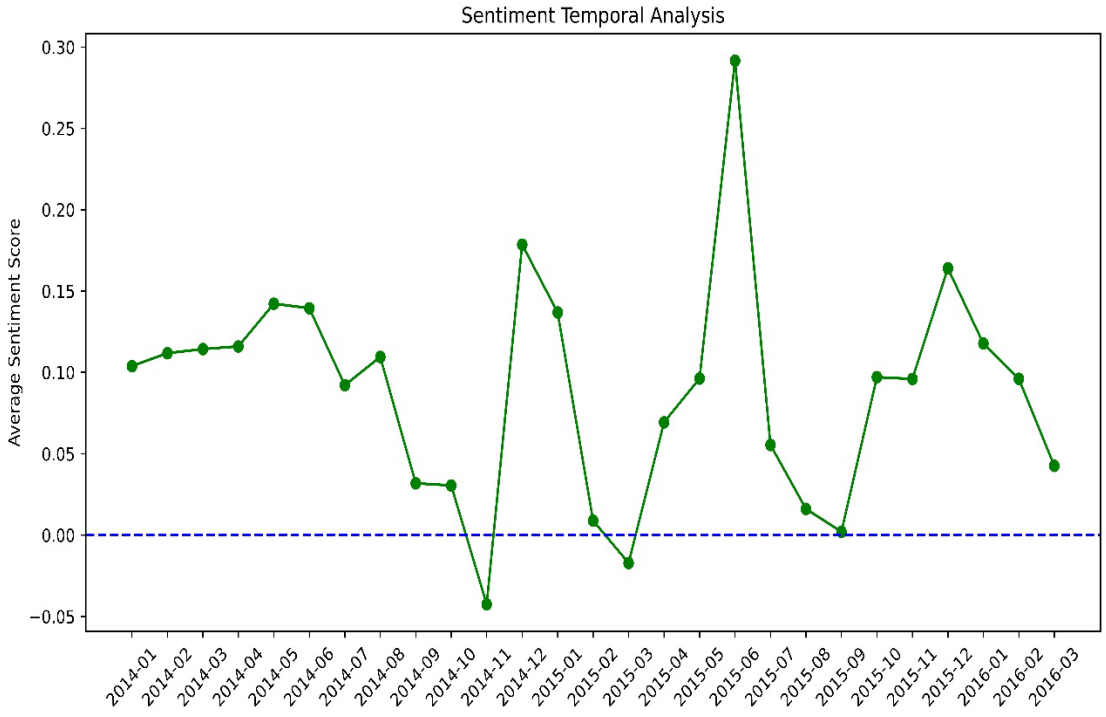

fig 4 Temporal Sentiment Analysis

The temporal sentiment analysis on the tweets related to Boeing company stocks is plotted in *fig 4*. The plot shows the monthly sentiment on the company which is the average of the daily sentiment for 30 days. Temporal sentiment analysis assist to gain the insights on the change of sentiments over the time. This kind of analysis provides the insights on possible situations and events that triggers the change of sentiment at a certain time. For example on 10th January 2015, Ethiopian Airlines Boeing 737-43QSF crashed in Accra-Kotoka, Ghana[7]. This led to the discussion on

this topic in social media which is generally negative, this dropping the sentiment score to negative. Similarly, at mid-2015s [Boeing Achieves Record Commercial Airplanes Deliveries][8], this led to the public sentiment gaining sharp rise in terms of the opinions and the confidence on the company.

This can be very useful to prepare, train and forecast the events in terms of stock market and investor's sentiment using the machine learning models. Trends over time, impacts of news and media, along with the event-driven changes can easily be identified by analysing the temporal sentiment trends.

Named Entity Recognition (NER)

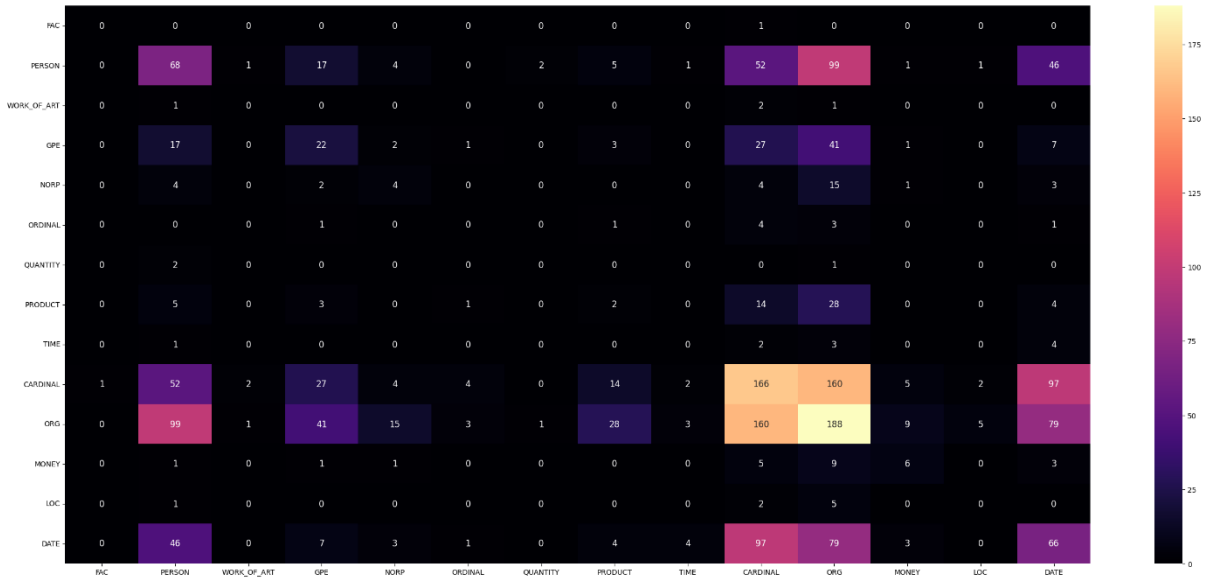

fig 5 NER heatmap

The heatmap on fig 5 shows the co-occurrence of different entity types in the data. Some of the entities like CARDINAL (numerical data that does not fall under another type), ORG (organizations), and PERSON (people's names) are frequently identified and sometimes co-occur with each other. The highest co-occurrence is between CARDINAL and ORG.

fig 6 NER Word Cloud

The word cloud from fig 6 visualizes the most frequent terms in the dataset, with “boeing” being highly prominent as expected due to the topic specific tweets. Other entities like “ge” (General Electrics, engine manufacturer), “bae systems” (Major defence maufacturer), “gt” (Goodyear Tire&Rubber, major tire manufacturer for the Boeing company), “msft” (Microsoft), “fb’ (Facebook), etc., are recognized by the algorithm. Thus the word cloud successfully depicts the entities mentioned in the tweets and they are highly related to each other showing the prominent players in the sphere in aerospace and technology.

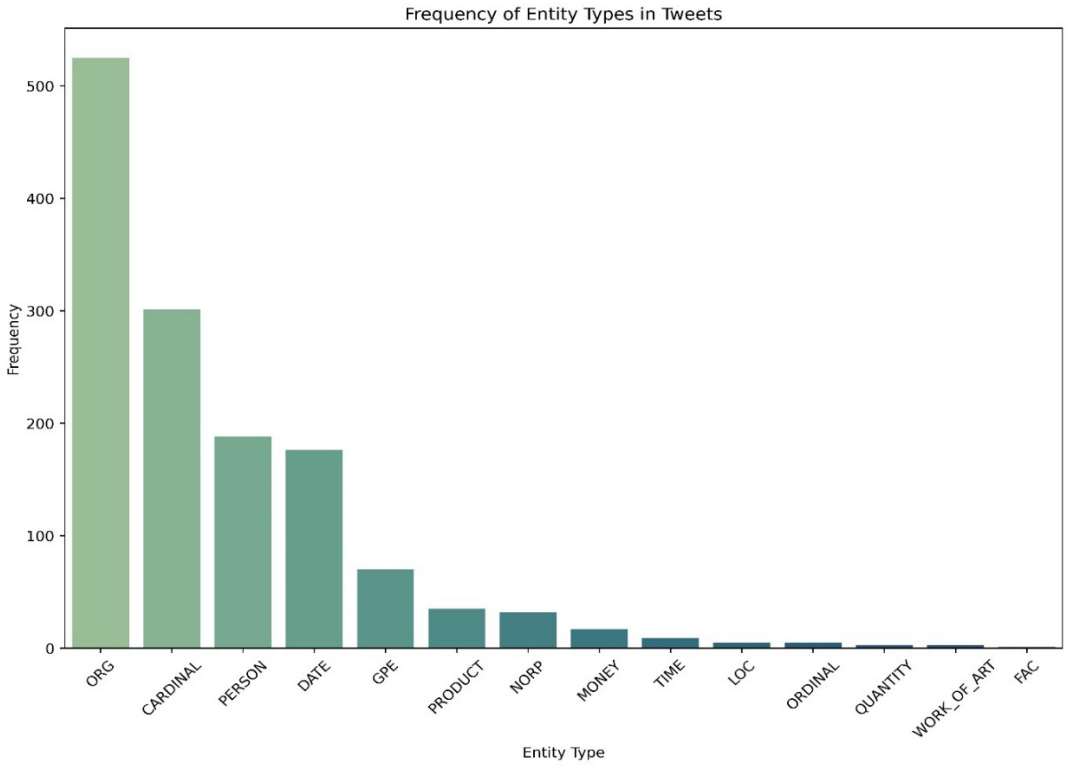

fig 7 NER entities type

Bar chart indicates the frequency of entity types in tweets. The ORG category is the most frequently occurring entity, followed by CARDINAL and PERSON. This suggests that organizations are a central topic of discussion in the dataset, which aligns with the word cloud's emphasis on company names like 'Boeing'.

Conclusion

In the analysis of Boeing company stock-related tweets, sentiment analysis using VADER revealed a normal distribution of sentiment scores, with a significant number of neutral tweets. Positive sentiments were associated with terms like "earnings," "trade," and "profit," while negative sentiments were linked to words such as "low," "beat," and "problem." Temporal sentiment analysis showcased monthly trends, illustrating how real-world events, like the Ethiopian Airlines crash, impacted sentiment. Named Entity Recognition (NER) identified key entities, including Boeing, General Electric, BAE Systems, and others, emphasizing their

centrality in discussions. The co-occurrence patterns highlighted the relationships between entity types.

The integration of sentiment and NER analysis provided a comprehensive understanding of sentiment dynamics and key entities influencing discussions. The results offer valuable insights for investors and stakeholders, enabling informed decision-making and proactive responses to market changes. The combination of temporal sentiment analysis and NER enhances the analysis's depth, making it a valuable tool for monitoring sentiment trends and identifying influential entities in the context of Boeing stock discussions.

References

1. https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=analytics-about-text-mining
2. Turney, P. D., & Littman, M. L. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *arXiv preprint cs/0212012*.
3. Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.
4. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.
5. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
6. Grishman, R., & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
7. https://www.1001crash.com/index-page-description-accident-Ethiopian_B737F-lg-2-crash-369-ethiopian-airlines-boeing-737-freighter-ghana-accra-kotoka.html
8. https://boeing.mediaroom.com/2016-01-07-Boeing-Achieves-Record-Commercial-Airplanes-Deliveries-in-2015