**Introduction**

Twitter, as of 2023, has approximately 530 million active monthly users, whose sentiments, when related to a particular topic, business or product, can be used to estimate the public perception of a brand, monitor the response to a new product or service launch, monitor a competitor’s activity, and so on (Hutt, 2023). Sentiment analysis – applied to a Twitter feed, or any other body or bodies of text – is the mining of text data to subtract subjective information from the data (i.e. the sentiment underlying it). Sentiment analysis is the most commonly applied and utilised form of text classification (Gupta, 2018).

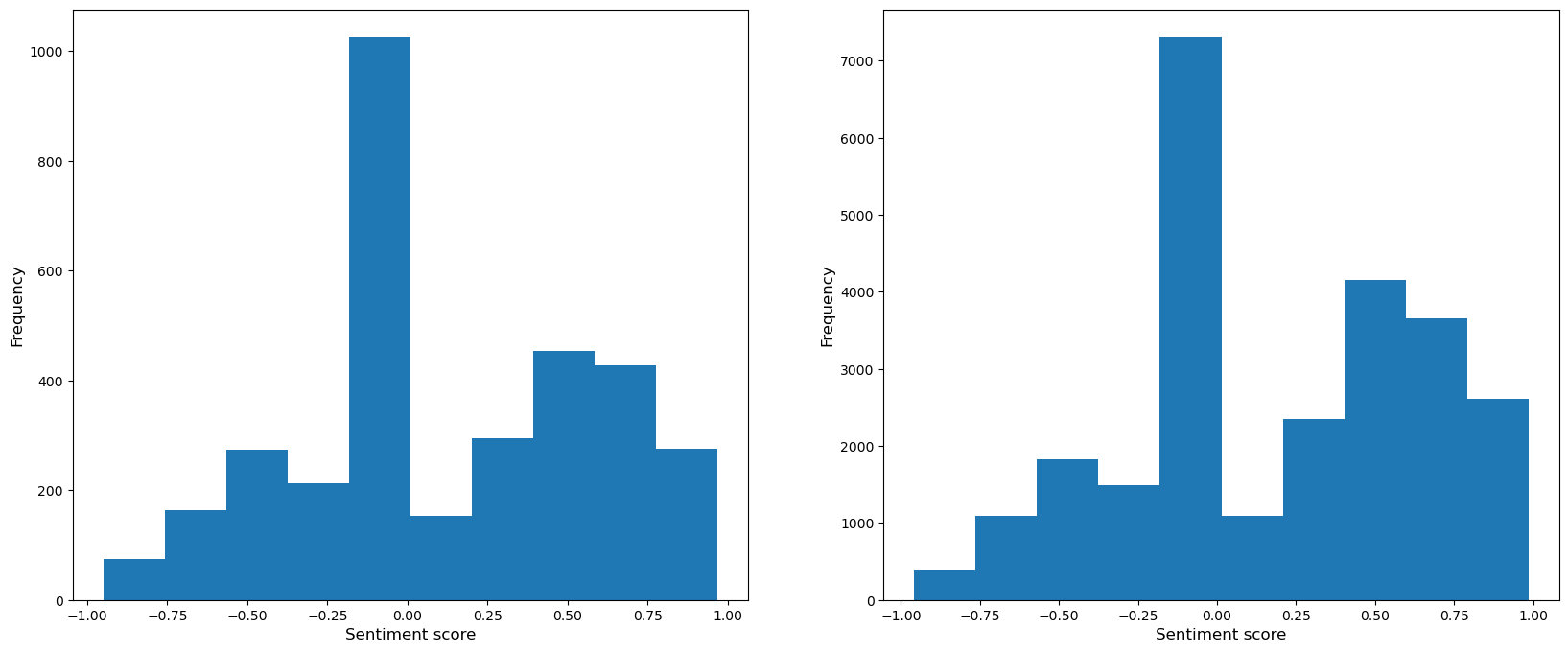
A time series dataset consisting of 6 million tweets were to be analysed for sentiment, and this sentiment forecasted for 3 time periods, 1 week, 1 month, and 3 months. There was a high proportion of days in the 81 day time series available for analysis that contained no tweets at all, so missing data imputation had to be completed first using an appropriate non-parametric method, then forecasts produced for the relevant time periods.

**Method**

The first step was to do some initial EDA to start to understand the content and shape of the data. With 6 million tweets, and 6 variables, there was no missing data, though there were also no queries associated with any of the tweets. Also, many of the dates at which the tweets were recorded were not unique, as the field only recorded to the nearest second, while milliseconds or even smaller denominations may have be required to distinguish tweet submission time. Finally, the username field was not thoroughly examined, because it was not going to be used in this analysis of tweet sentiment over time, but it was found that there were a number of users who regular tweeted. It was also found that the tweet ID was, unexpectedly, not unique. This was because some of the tweet entries were duplicates. The mapper was used to remove these duplicated values by selecting only unique tweet IDs, leaving 1,598,315 unique tweets. This course of action was taken because it was assumed that the tweet duplication was a mistake. If these tweets had been posted twice deliberately - maybe for emphasis - they arguably should have been retained, but in that case the tweet ID would have been unique, and they were not.

The MapReduce program used took the raw data (saved to the Hadoop HDFS as shown in the Appendix) and returned only the date and punctuationless tweet. Screenshots of the MapReduce program running are included in the Appendix. A complication was that the tweets themselves often contained commas, and the mapper – programmed to split the data on a comma - was truncating those tweets with commas. To resolve this, a for loop in the mapper acted on any line longer than expected (i.e. any line that had been split into more parts due to the presence of a comma in the tweet) and concatenated the tweet segments together to form a single output tweet. The mapper returned the tweet number, the date, and the un-edited (other than removed commas) tweet. The reducer took this output and stripped all punctuation from it (i.e. removed all non-alpha characters), then returned the tweet number, the date, and the tweet, and output them to a comma separated file. The reason that the tweet letters were not all changed to lowercase was because the VADER model used to determine the sentiment distinguishes between capital and lowercase (Beri, 2020), and the case can impact the sentiment. VADER also assigns meaning to punctuation, so it may be advisable to leave the punctuation, or only remove aspects of it, such as ellipses or tildes, and leave more potentially meaningful ones such as exclamation marks, or question marks. However, punctuation was removed here to demonstrate a reducer programme. Stopwords, however, were not, to avoid removing context, especially as VADER can understand that adverbs, such as ‘not’, can change the sentiment of a word (e.g. ‘happy’, vs ‘not happy’) (Beri, 2020). The reducer output was used for sentiment analysis (after renaming to add an extension, .txt, to allow it to be read by pandas in Python).

The VADER model scores the sentiment of each word in a piece of text, and sums them to determine the sentiment of the text as a whole (Beri, 2020). Capitalisation and punctuation are considered by the model, for example, the word ‘Happy’ gives a score of 0.5719, while ‘Happy!’ gives 0.6114. It is part of the NLTK package, one of the most commonly used Python natural language processors, and is simple and fast (DiBattista, 2021). As a commonly used, reliable, and easy-to-use package, it was selected for use here. A score was calculated for each tweet, then the daily median sentiment score was determined. A daily median was selected as the unit of time, as it was considered that an hourly median may invite inherent bias (for example, people tweeting as soon as they wake up, or in the evenings, in different parts of the world may be reasonably expected to be more negative in areas and regions affected by war, for example). Also, a daily median would be less likely to be impacted by missing values, and would usually be expected to contain a relatively large dataset, with a spread of sentiment scores approximating a statistical distribution that could be modelled. For example, the 4th of June 2009 saw 3,354 tweets, and the 10th of May 2009 saw 25,984 tweets. As shown in Figure 1 both sentiment score distributions appear bimodal (non-parametric) (hence the use of the median as the measure of central tendency).

Figure 1: Two daily sentiment score distributions showing non-parametric nature of the data.

Tests for stationarity showed that the time series could be considered stationary (augmented Dickey–Fuller statistic: -5.26, DF p-value: 6.48 x 10-6; KPSS statistic: 0.55, KPSS p-value: 0.03), but additive decomposition (on the data after missing values were imputed) indicates seasonality, as shown in Figure 2. A significant proportion of the data - 69% when considering daily medians – consisted of missing values. These needed to be dealt with before any forecasting was attempted. The data is assumed to be MCAR, as there is no known reason why there would be no tweets at all on the missing dates, and there is no common factor amongst the missing data (Yang and Chiang, 2020).

Not only was the sentiment score of all tweets in a day non-parametric, but so was the daily median score distribution for the 81 days available, as shown in Figure 3, and by the results of a [Shapiro–Wilk tes](https://en.wikipedia.org/wiki/Shapiro–Wilk_test)t (statistic=0.71, p-value=2.51 x 10-8). As the data was non-parametric, non-parametric imputation, such as decision-tree and KNN methods, were most appropriate (Bertsimas, Pawlowski and Zhuo, 2018).

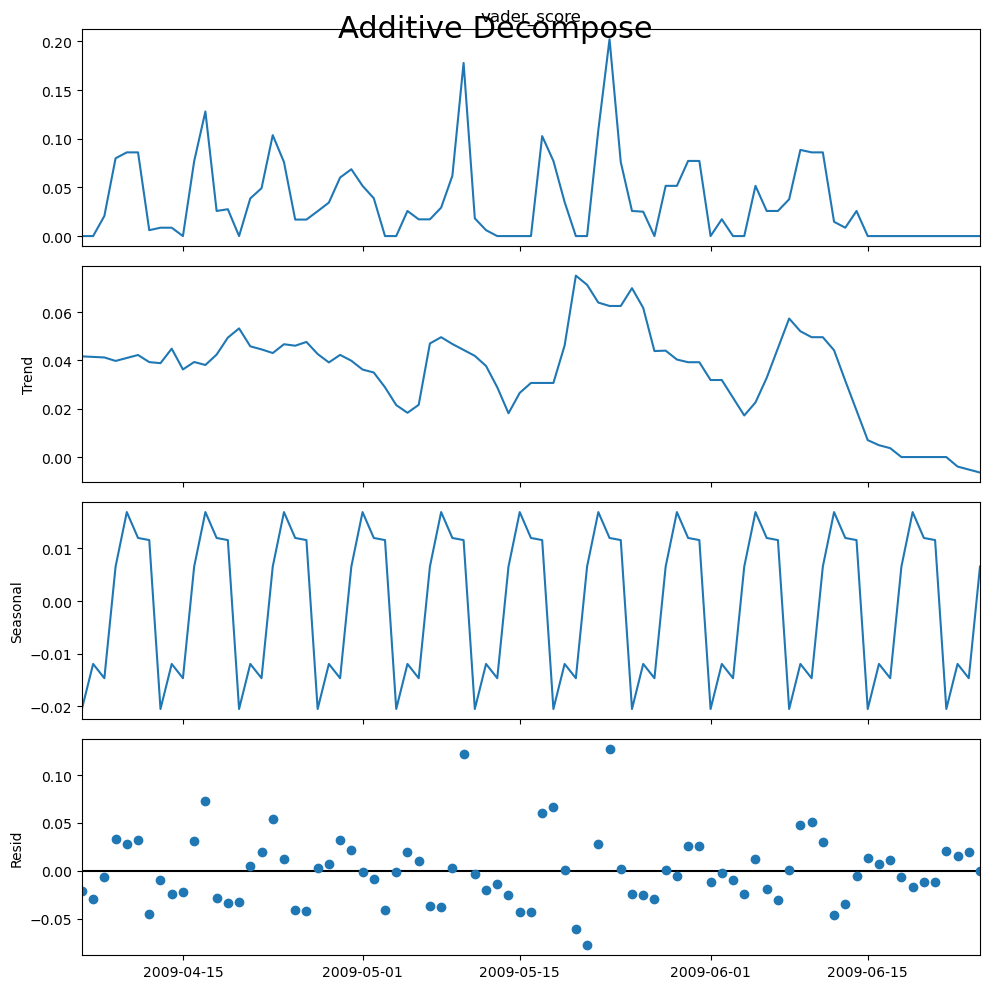
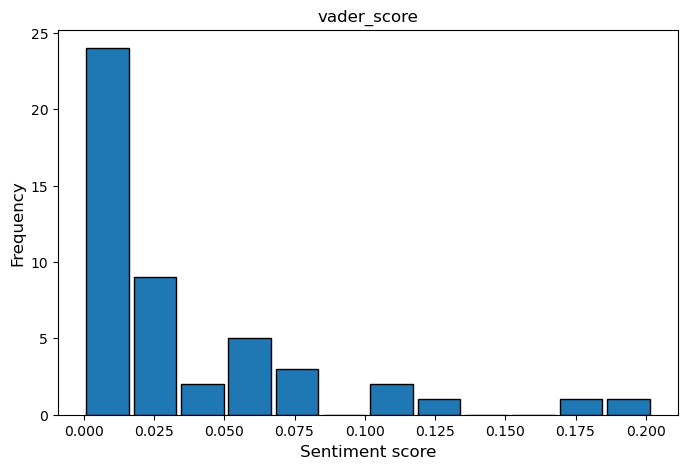
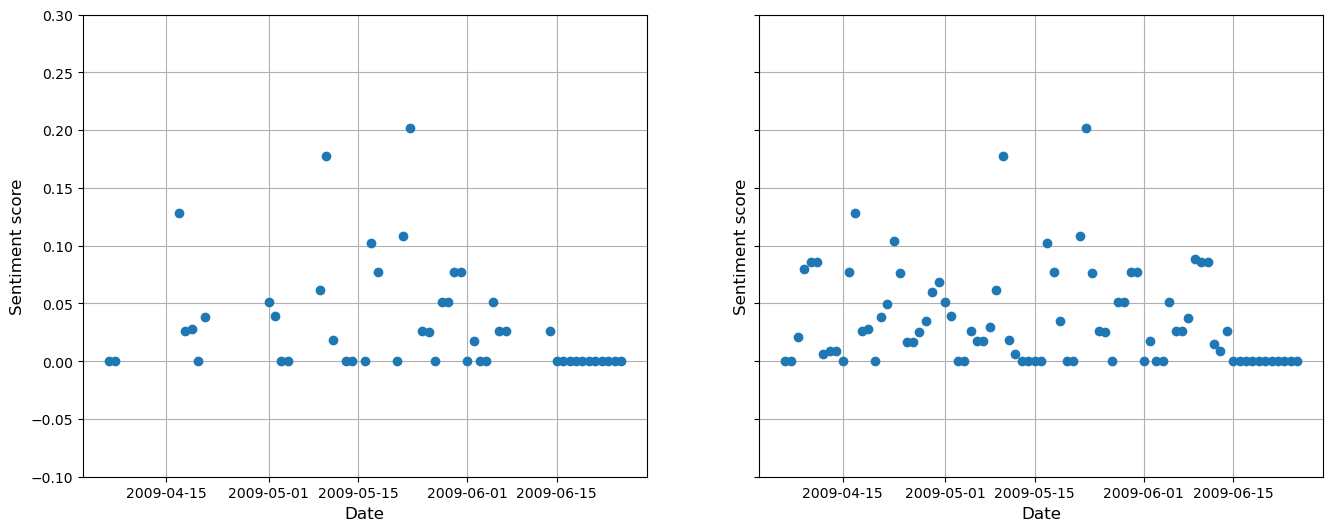


Figure 2: Additive decomposition on sentiment score time series data after missing values were imputed.

Figure 3: Histogram of the daily median sentiment score

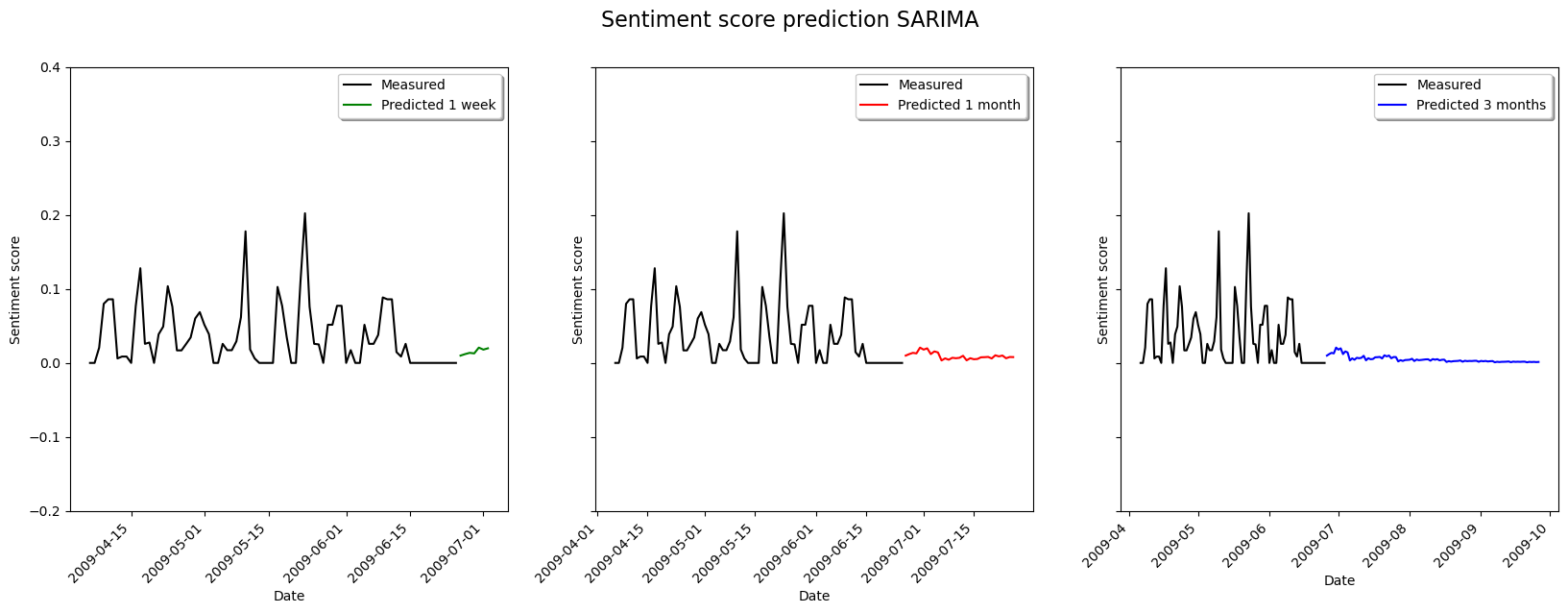
KNN imputation was used here, with 3 nearest neighbours selected. A comparison of the results with varying numbers of nearest neighbours can be seen on the dashboard. The day of the month and the month data were extracted for each data point, to allow the KNN imputer to work with these two aspects of the date independently. Ideally the data would have been split into test and train, and the parameters of the imputation selected by hyperparameter tuning, but there were no aspects of the time series long enough, and without missing values, to allow this. The imputed values in this case must be considered with some caution fbecause of this, and because of the high percentage of them in the original dataset (which is potentially why there is still evidence of seasonality in the residuals in Figure 2). The time series after missing values were filled is shown in Figure 4.

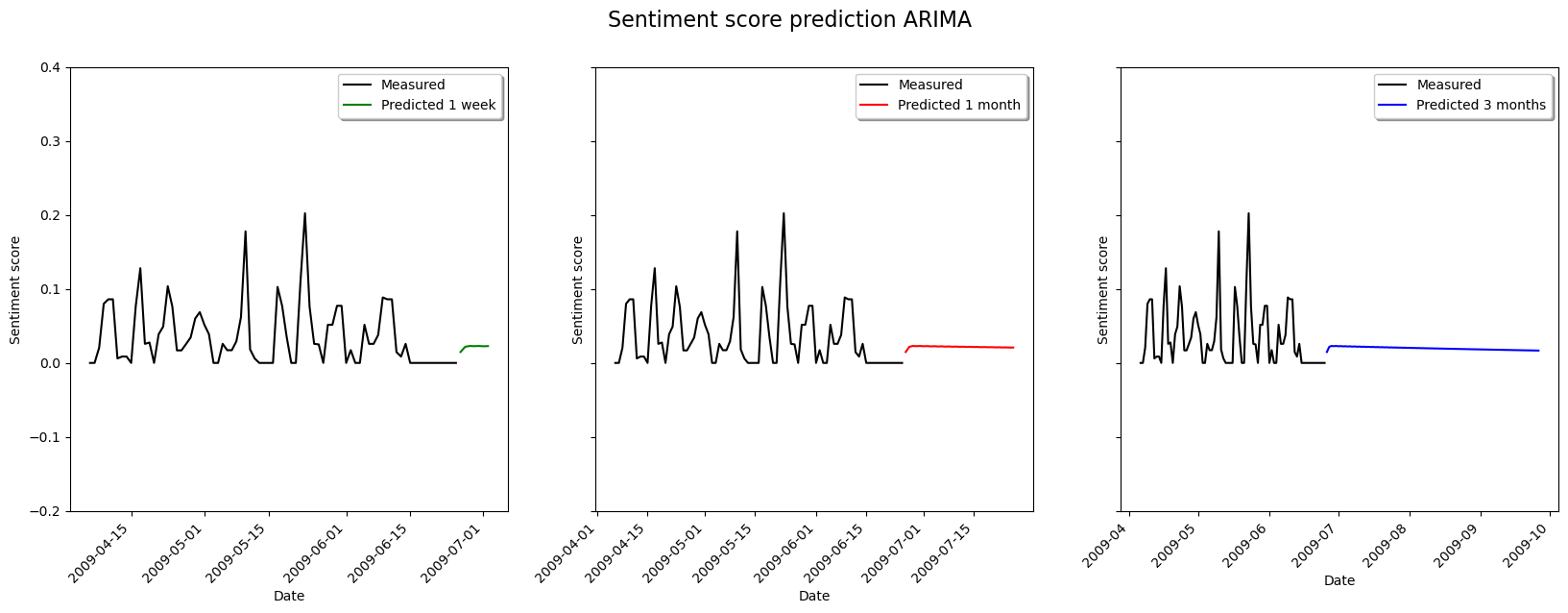
Figure 4: Before and after imputation (left, showing missing data) and right (showing imputed data)

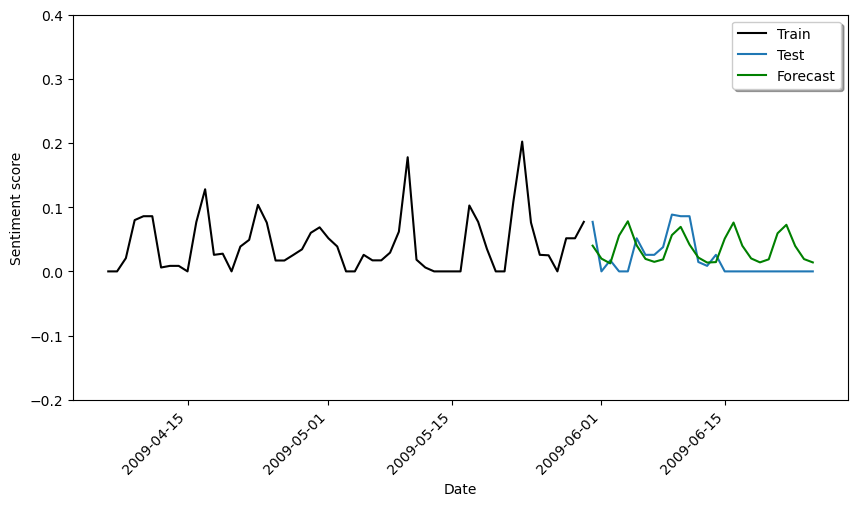
This data with imputed values was imported to two databases, Cassandra and HBase, as shown in the Appendix. These two databases were also load tested using ycsb and the default ycsb time series load tester, tsworkloada. The commands used to run these are in the appendix. To use ycsb with Cassandra a change was needed to conf/cassandra.yaml, namely to uncomment two commitlog\_sync lines (Brianfrankcooper, n.d.), and to run the Cassandra load test in ycsb a ycsb keyspace had to be created, as shown in the appendix, and similarly a usertable had to be created in hbase. While the HBase test worked, the Cassandra one did not fully run, with a latency and INSERT-ERROR error experienced. This is presumably an error in inserting into the Cassandra keyspace, but the issue was not found, so the results cannot be accurtely compared here. Hbase’s runtime was 308,878 ms, with almost 4,774 operations / second, while Cassandra’s results reported a runtime of only 2,745 ms, but with 0 operations / second, as the insert was not working as expected. Both results outputs are in the appendix.

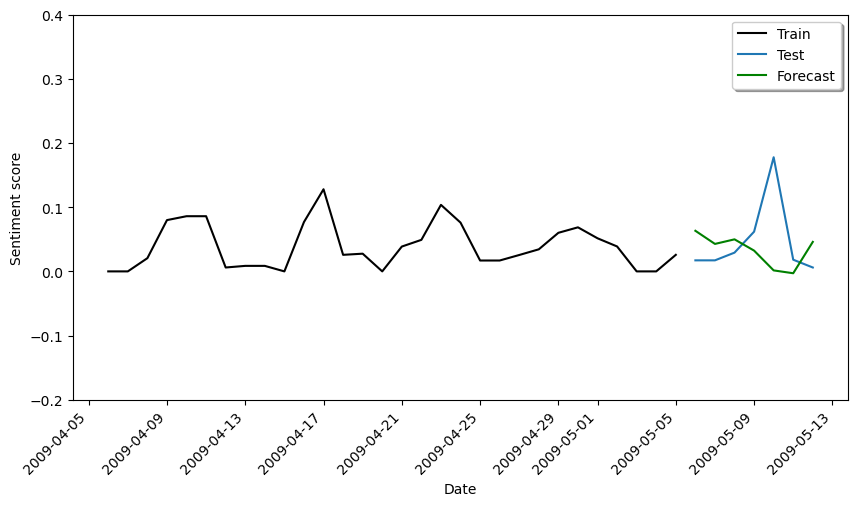
Finally, a time-series forecast was generated using ARIMA and, more appropriately, SARIMA. ARIMA does not account for seasonality, which this time series possessed, as per Figure 2, while SARIMA does. Autocorelation, partial autocorrelation and differencing were used to find the appropriate values of p, d and q, and the seasonal term in SARIMA was selected by examining the seasonal aspect of the additive decomposition in Figure 2. As the data was trend free, d (the number of non-seasonal differences needed for stationarity) was set to 0. The autoregressive term, p, was set to 3 as the number of lags that crossed the significance threshold in a partial-autocorrelation plot of the data, and q – the number of lagged forecast errors - was also set to 3 from an autocorrelation plot of the data (Verma, 2022; people.duke.edu. (n.d.)). The results of the 1 week, 1 month, and 3 month SARIMA forecasts are shown in Figure 5. For comparison, the ARIMA model, which does not account for seasonality, is shown in Figure 6. This latter model forecasts the same median sentiment (just above 0) for each day into the future, with a slight downward trend. In contrast, the SARIMA model peters out to a median sentiment of 0 toward the end of the 3 month period.

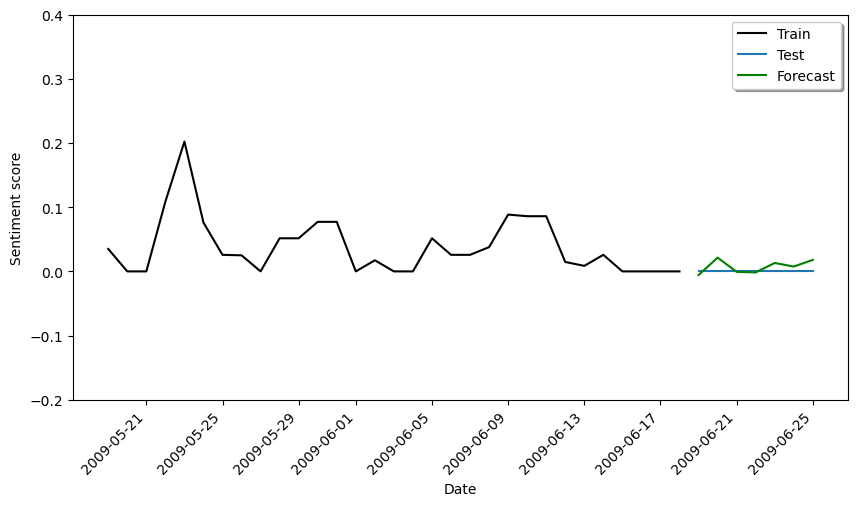
To test the reliability of these SARIMA forecasts, and having selected the appropriate p, d, q, and s values, the dataset was split into test and train to quantify the length of reliable forecast that could be achieved from the data at hand. The forecast accuracy was always impacted by the drop in sentiment over the last 2 weeks or so, as there was no precedent for this is the data available. Splitting the data into 1/3 test, 2/3 train (this was close to a 2 month / 1 month split – 26 days test and 55 days train), the forecast was comparatively good (RMSE: 0.038, Figure 7), though forecasting a week of sentiment from a month of data was less so at the start (train data from 6th April – 5th May, RMSE: 0.073, Figure 8) but surprisingly more so towards the end of the series (train data 19th May – 18th June, RMSE: 0.012, Figure 9), despite the drop in sentiment towards the end. Forecasting nearly 2 months (55 days) from nearly 1 month (26 days) was also not very reliable (RMSE: 0.050, Figure 10).

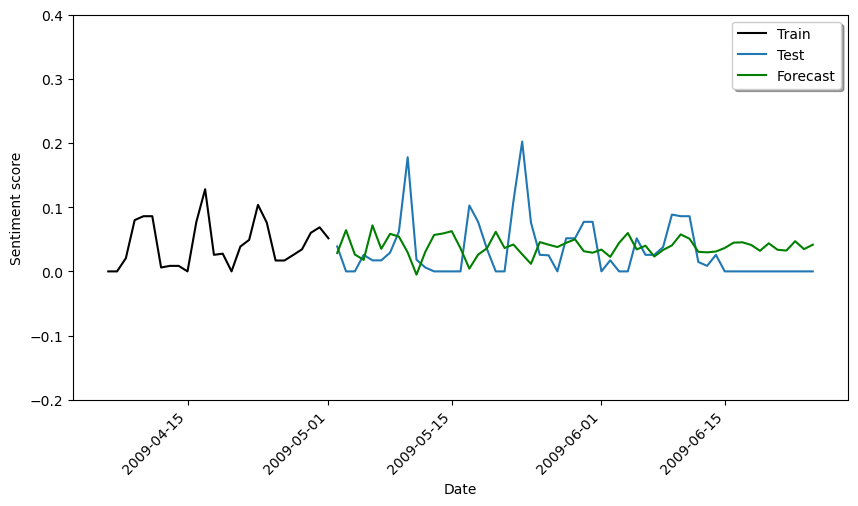
Figure 5: SARIMA forecasts of sentiment score over 1 week, 1 month, and 3 month periods.

Figure 6: ARIMA forecasts of sentiment score over 1 week, 1 month, and 3 month periods.

Figure 7: Forecasting 26 days based on 55 days train data.

Figure 8: Forecasting 7 days based on 1 month train data, at the start of the available series.

Figure 9: Forecasting 7 days based on 1 month train data, at the end of the available series.

Figure 10: Forecasting 55 days based on 26 days train data.

**Results and Discussion**

There were two main results of the analysis, the validity of the imputed missing values, and the validity of the forecasted sentiment. Using KNN with 3 nearest neighbours allowed the imputation model to select values with similar dates and account for the seasonality of the data. A much longer data would have been need to do this more accurately, and to allow the use of train and test datasets to train the model. The main impediment to that here was the lack of any significant length of time without missing data to train on. An increase in the number of nearest neighbours – as shown in the Dashboard – just acted to smooth the resulting imputed time series, but did not significantly impact the periodicity / seasonality of approx. 6 days in length. Instead, the amplitude of the peaks in sentiment were reduced. This gives some evidence that the seasonality is an inherent feature of the data’s sentiment, rather than an artifact of the imputation.

In terms of the forecast, the model was designed – by setting the number of non-seasonal differences needed for stationarity to zero – not to model any trend, and only to forecast for seasonality. This was complicated by the fact that the median sentiment appeared to drop significantly, and the seasonality to disappear, over the last few weeks of the available data. The forecast initially emulates the seasonal pattern for the first week or so – approximately corresponding to the period of the seasonality, but then tapers gradually towards zero with increasing time. Generally, the longer the time series available, the better the quality of the forecast made using is – though the impact of longer team series starts to peter out when adding even more time to already long time series. There are time series where the optimal length for forecast is actually quite short (even down to 1 month can be sufficient), and that is a result of underlying patterns changing frequently (Svolba, 2022). This may even apply to tweet sentiment data, though it is not possible to test that for a three month forecast, as there was less than three months available.

Due to the unprecedented sentiment drop towards the end of the time series available it was not reasonable to expect that a model trained on the otherwise seasonal data would be able to forecast the sudden decline in sentiment. The reason for this drop is not known, as the context of the tweets is not known, but something similar could conceivably happen in response to a piece of negative press about a company, or due to a cessation of advertising around a product. The reason that the forecast started to improve when only forecasting a week towards the end of the series is likely because sentiment had already started to drop as part at the end of the training period. The results of testing the SARIMA model shows that longer forecasts from relatively little data is not reliable, though forecasting shorter time periods also is not always reliable – likely due to the variability not fully captured by the seasonality, as shown by the residuals in Figure 2.

While sentiment analysis is undoubtedly useful, and is very widely used, there are other – potentially equally informative – analyses that may be applied to data to extract subjective meaning from it, such as intent analysis – extracting the authors intention, e.g. marketing or querying - and contextualised semantic analysis - segregating messages into groups such as product areas, or staff vs product feedback, etc., and carrying out sentiment analysis or intent analysis on that (Gupta, 2018).

There was no known theme or subject area in particular to give context to the body of the tweets, so the reduction in sentiment to a median of zero towards the end of the 81 day period in June 2009 cannot be commented on without knowing what may have been happening in the wider sphere. Throughout the time period there was an approximately 6/7 day periodicity to the sentiment score values, potentially indicating a response to a periodic event, such as an advertising run, or maybe something that was influenced by the occurrence of a weekend - a particular location, or alcoholic beverage maybe that is most popular / talked about at the weekend. Had a subject area been known, collocations could have been used to extract or search for meaningful word pairings. With a longer dataset, the changing sentiment of individual users could have been analysed with time to get a snapshot of not just changing sentiments in general, but also individual people’s feelings on a subject.

**Conclusion**

An 81 day twitter feed snapshot was analysed for sentiment, and, after having a sentiment score applied using the VADER tool from NLTK. Missing values were imputed using a 3-nearest neighbour KNN method, and the dataset was found to have a seasonality element to the time series, but no trend. SARIMA was therefore an appropriate forecasting tool, but was not found to be very reliable due to the varying nature of the seasonality, and the drop in sentiment towards the end of the time-series.

**References**

Beri, A. (2020). *SENTIMENTAL ANALYSIS USING VADER*. [online] Medium. Available at: <https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664>.

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2018). From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research*, 18, pp.1-39.

DiBattista, J. (2021). *The Best Python Sentiment Analysis Package (+1 Huge Common Mistake)*. [online] Medium. Available at: <https://towardsdatascience.com/the-best-python-sentiment-analysis-package-1-huge-common-mistake-d6da9ad6cdeb>.

Brianfrankcooper, GitHub. (n.d.). *Improvements for HBase and Cassandra clients by jbellis · Pull Request #98 · brianfrankcooper/YCSB*. [online] Available at: https://github.com/brianfrankcooper/YCSB/pull/98/files [Accessed 7 Nov. 2023].

Gupta, S. (2018). *Sentiment Analysis: Concept, Analysis and Applications*. [online] Towards Data Science. Available at: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

‌

Hutt, S. (2023). 7 Advantages Of Twitter Sentiment Analysis. [online] Available at: https://ecommercefastlane.com/7-advantages-of-twitter-sentiment-analysis/ [Accessed 7 Nov. 2023].

people.duke.edu. (n.d.). Introduction to ARIMA models. [online] Available at: https://people.duke.edu/~rnau/411arim.htm#:~:text=A%20nonseasonal%20ARIMA%20model%20is.

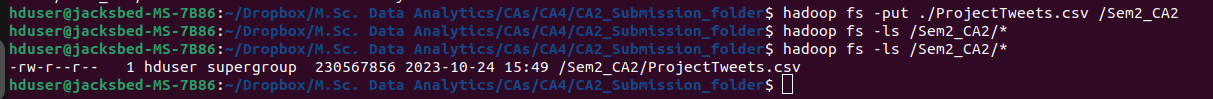
Svolba, G. (2022). Determining the best length of the history of your timeseries data for timeseries forecasting. [online] MLearning.ai. Available at: https://medium.com/mlearning-ai/determining-the-best-length-of-the-history-of-your-timeseries-data-for-timeseries-forecasting-f8600a3c086#:~:text=than%202.5%20years.- [Accessed 8 Nov. 2023].

Verma, Y. (2022). Quick way to find p, d and q values for ARIMA. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/>.

Yang, L. and Chiang, J.A. (2020) ‘Use case and performance analyses for missing data imputation methods in big data analytics’, Proceedings of 2020 6th International Conference on Computing and Data Engineering [Preprint]. Doi:10.1145/3379247.3379270.

**Appendix**

**Big data operations carried out outside of Jupyter notebook**

Figure 11: Saving raw data to Hadoop HDFS before MapReduce carried out.

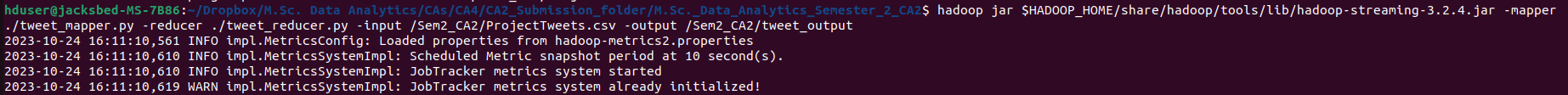
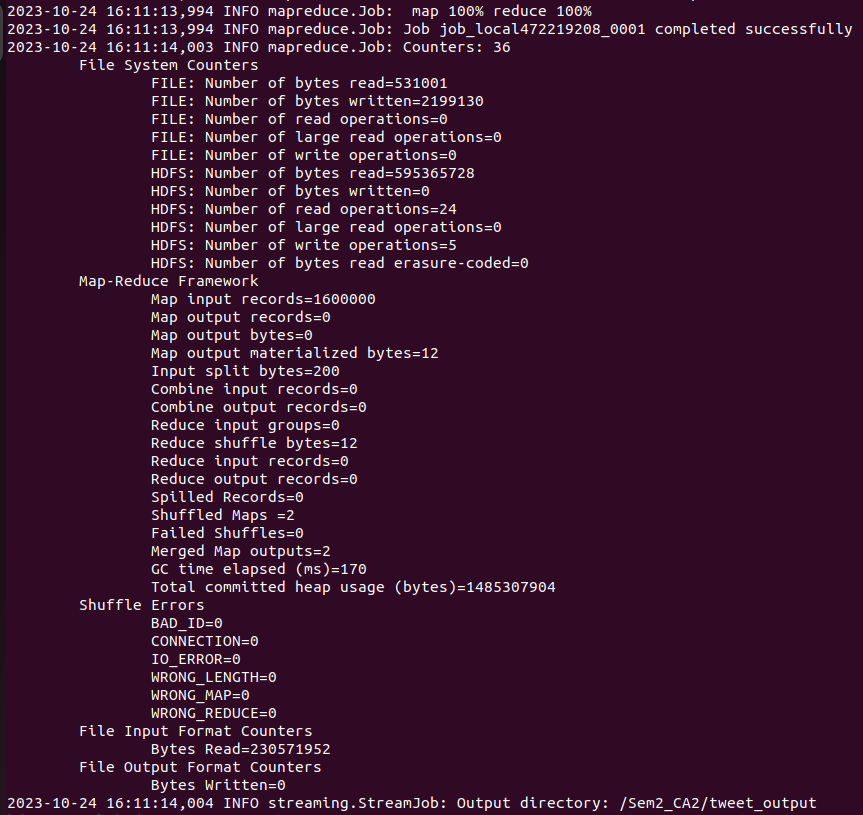
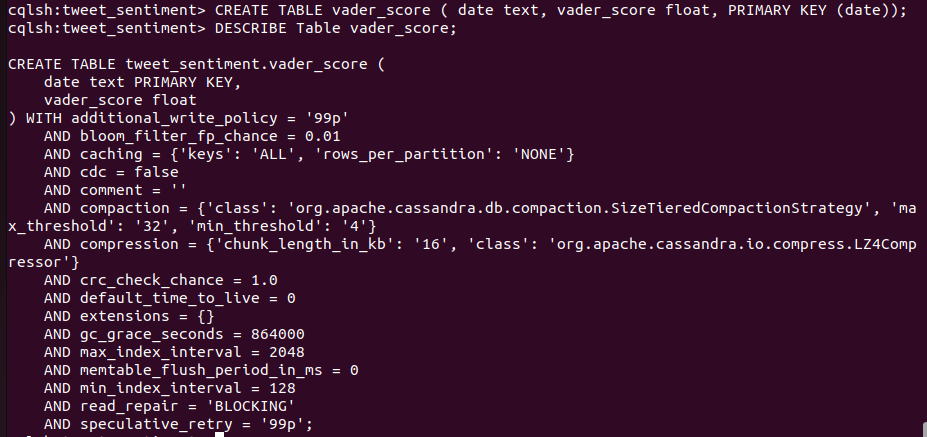
Figure 12: MapReduce program in operation

Figure 13: Mapper section complete

Figure 14: MapReduce complete

Figure 15: Creating a table for the sentiment score daily median data in Cassandra, with the date as the primary key.

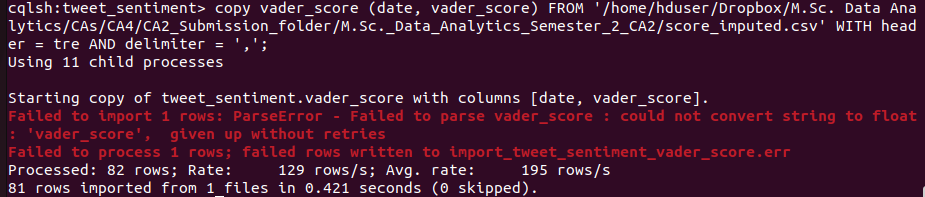
Figure 16: Copying the median daily sentiment score data to Cassandra

Figure 17: Selecting data from the sentiment score table in Cassandra

Figure 18: Creating a table for the sentiment score daily median data in HBase

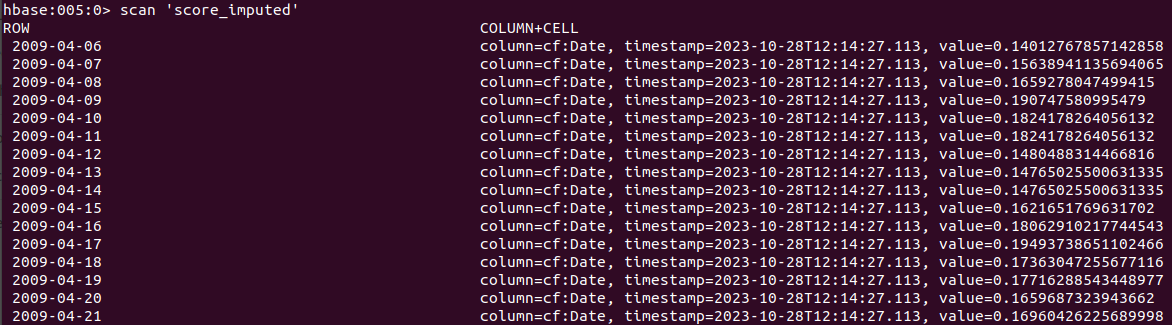
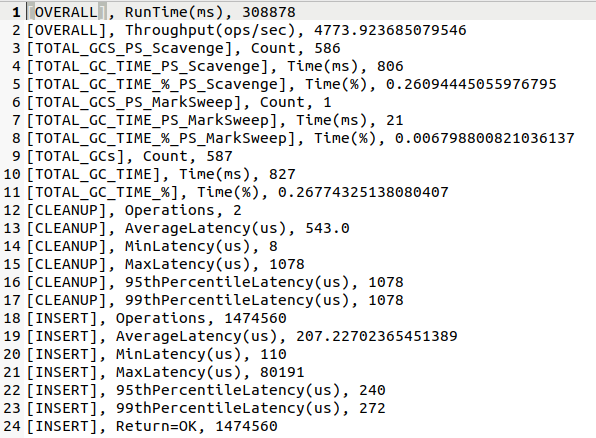
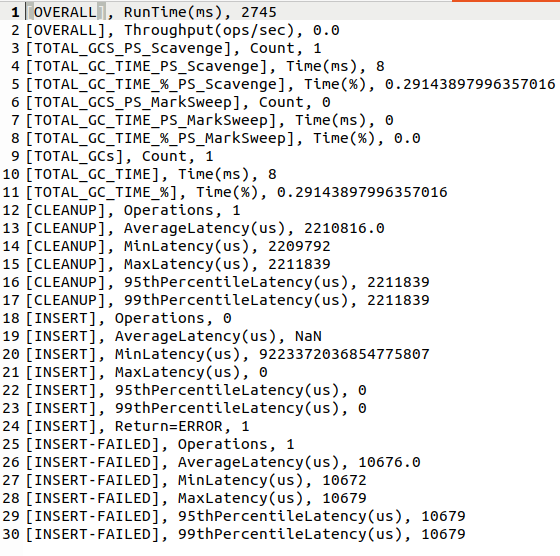
Figure 19: Scanning the HBase table for the sentiment score daily median data.

Figure 20: ycsb keyspace creation for the Cassandra load test

Figure 21: Running a ycsb load test on Cassandra cql

Figure 22: HBase ycsb results

Figure 23: Cassandra ycsb (failed test) results