# A Case Study for the Benefits of Cluster Analysis of Social Media Data and Retailer Sales for Twitter and A UK Based Department Store

Tommy Hamm[1, 2], Enda Fallon[1], Sheila Fallon[1], Paul Connolly[2], Kieran Flanagan[2]

[1]Software Research Institute, Athlone Institute of Technology, Athlone, Ireland
thamm@ait.ie, efallon@ait.ie, sfallon@ait.ie

[2]The NPD Group, Inc Athlone, Ireland
paul.connolly@npd.com, kieren.flanagan@npd.com

***Abstract*** **- Due to the continuous growth of online interaction, social media is becoming increasingly useful in understanding trends in human behavior both locally and globally. On average there are approximately 6,000 tweets posted on Twitter every second, equating to approximately 500 million tweets per day. This wealth of information shared publicly can be hugely beneficial in gaining insights into reactions and implications caused by social, environmental, or financial events. The information has the potential to be particularly useful to retailers in terms of market research and sales forecasting when used along with some of the latest data analysis and Artificial Intelligence (AI) tools. The goal of this study is to utilize data from the Twitter platform, shared by the public, to extract what benefits and insights can be gained by analyzing the correlation between external KPIs, extracted from non-UK based geographical social media data, and sales recorded in a UK based luxury retailer at the corresponding time.**

***Keywords - social media, trends, Twitter, data analysis, AI, KPIs, sales forecasting***

## I. INTRODUCTION

Customer buying behavior is constantly changing and new business possibilities and risks emerge as customer's preferences shift and the retail landscape develops. Modern-day AI tools are becoming an increasingly crucial part of leveraging data from across e-commerce platforms, department stores, and distribution centers to help retailers become more agile and deliver a more personalized and convenient shopping experience to their customers.

Today, with the data that is accessible, it's possible to gain a good understanding of consumer e-commerce platform interactions. Analyzing user's online shopping experiences, looking at what products consumers are clicking on, and what they're buying at the end of their online shop is an example of the kind of data and visibility available. This visibility is evidently less clear in the actual stores. Finding a way to gain intelligence surrounding the behavior of consumers in the store and combining the knowledge of customers' behavior online could provide full 360° visibility of consumer's behavior.

This initial study takes a regional approach of analyzing customer's behavior by utilizing social media data retrieved. The study analyses Twitter activity, based on a UK luxury department store, to identify any correlations between external KPI's extracted from the Twitter data and sales recorded for the retailer.

Social Media has the potential to discover beneficial insights on human emotions and behavior at a personal, regional, and global scale. Having the capability to scrape historical Tweets and continuously monitor the latest Tweets can provide benefits in analyzing situations and the ever-changing opinions of the public on a given topic, this use case being a given retailer. To be able to gain these useful insights from publicly shared Twitter posts, an application was developed to retrieve, and store filtered tweets on a given topic, both historically and in real-time. This was utilized to collect Tweets in relation to a given UK retailer over a 2-year period.

The NPD Group, Inc provides data, industry experience, and prescriptive analytics to customers to help them expand their businesses in a changing environment. The world's most successful businesses rely on NPD to help them assess, forecast, and optimize sales performance. A tool that NPD utilize and have made accessible for the purpose of this study is one that allows feeds of anomaly detection in point of sales (POS) data. This information and the social media data retrieved, both based around the 2 Year time scale, for the given UK retailer is the bassline data source for the study. The aim of the study is to hypothesize if external KPIs extracted from social medial data might be useful in enhancing current AI and Analysis tools.

## II. RELATED WORK

Researchers are increasingly studying ways to try and present various ideas which involves the use of social media data. This very accessible data has been utilized for ideas such as, a public perception monitor, a public health surveillance data source, and on some instances has shown the benefits of extracting the diverse amount of information from various social media platforms to provide insights on the direction of a public election or movements on the stock market.

Recent research has shown that tweets can be used to predict various large-scale events like elections and national revolutions. The essential hypothesis is that the location, timing, and content of tweets are informative regarding future events. A paper written in 2014 from University of Virginia asked the question "can tweets, posted by residents in a major U.S. city, be used to predict local criminal activity?" [1] The paper showed that the addition of Twitter-derived features improves prediction performance for 19 of 25 crime types and does so substantially for certain surveillance ranges. The results indicated potential gains for criminal justice decision makers: better crime predictions should improve the allocation of scarce resources such as police patrols and officer time, leading to a reduction in wasted effort and decrease in crime response times [1].

Since January 2020, there have been a growing number of papers that analyze Twitter activity during the current global pandemic that is COVID-19 [2-6]. One Researcher analyzed the frequency of 22 different keywords such as "Coronavirus", "Corona", "CDC", "Wuhan", "Sinophobia", and "Covid-19" analyzed across 50 million tweets from January 22, 2020 to March 16, 2020 [7]. Another paper published an analysis of topics for English-language tweets from March 10-29, 2020 and analyzed distribution of languages and propagation of myths to understand perception of public policy from the Twitter information to model information and misinformation spread. [8].

Currently in the retail industry there is increasing activity around the use of artificial intelligence, one use case looks at leveraging computer vison and algorithms that integrate into point of sales systems to detect and prevent theft in real time.

A recent study looks at a brief analysis of retail customer's consumptions experience under the background of AI. The paper concluded that to make AI technology better serve the retail industry, more attention should be paid to the formulation of perfect laws and regulations and industry system while further researching and developing related technologies. [9] Another paper looked at an AI-based customer behavior analytics system, this involved implementing a deep learning-based in-store traffic monitoring system for evaluation of retail performance [10].

There is a vast amount of research carried out on the use of AI in the retail industry and separately in AI and social media. However, there is very little carried out in combining both these areas along with AI tools and this paper looks at the benefits this could potentially achieve. As of recent, one paper did research the role of AI in social media marketing. The objectives of the paper was to study the scope of artificial intelligence in marketing, to study the pros and cons associated with the use of AI in marketing and to study the attitude of marketing managers towards AI. The paper concluded that, AI is revolutionizing industries with its vast applications and helping in solving complex problems. There is no proof required to accept the growth of AI in marketing. With the help of AI, marketers are able to identify potential customers, create content, and follow the leads. Incorporating AI can assist marketers, individuals, and advertising agencies in making social media marketing more efficient.

## III. SYSTEM DESIGN

### *A. Data Retrieval and Storage*

A RESTful web service was developed to interact with the Twitter API to retrieve and store Tweets. A RESTful web service exposes a set of resources that identify the targets of the interaction with its clients. Resources are identified by URIs, which provide a global addressing space for resource and service discovery. To manage the retrieval and storage of Tweets a GET and POST endpoint were developed described in the Table below.

TABLE I. RESTFUL API ENDPOINTS

| Method | URI | Description |
|---|---|---|
| GET | api/v1/tweets | An endpoint to retrieve all tweets stored in the Postgres Database. |
| POST | api/v1/tweets | An endpoint that uses the twitter4j library to search for filtered based tweets to be Parsed and persisted to the Postgres Database. |

Figure 1 below represents the Architecture of the application developed to retrieve and store Tweets. The Spring framework was used to develop an API Layer using Java EE for the GET and POST endpoints described above, Postgres was used to develop a Data Access Layer for parsing the JSON response of Tweets from the Twitter API and storing them in a database for retrieval when required.

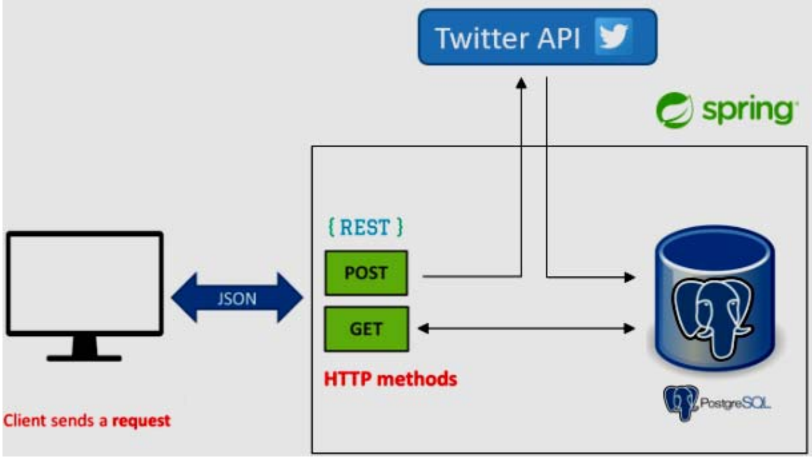

Figure 1. Application Architecture

### B. Data Sources

Table II represents the features that the Twitter API exposes for the application developed to retrieve and Table III represents the information accessible from the tool made available from The NPD Group.

TABLE II. TWITTER DATA

| Feature | Description | Sample |
|---|---|---|
| lang | The language the Tweet was posted in | en |
| source | The platform in which the tweet was posted | Twitter Web App |
| context | The background on which the Tweet was primarily about | Retail |
| Created-at | The time and date the Tweet was posted | 2020-11-19 T01:20:32 |
| text | The content of the Tweet | Hello! |
| country | The Country of the user that posted the Tweet. | Ireland |

### C. Data Visualisation

Visuals have been used for Millenia as a method of communication of ideas and concepts and in recent years technologies in relation to data visualization have improved immensely. Matplotlib, a comprehensive library for creating static, animated, and interactive visualizations in Python was used to produce some visuals for the insights discovered. The x axis in the visuals will represent a time scale over a given period and the y axis will represent a numerical value. The time series plot will visualize the volume of tweets against the number of sales in the given retailer, for the case study, at corresponding times.

TABLE III. NPD DATA

| Feature | Description | Sample |
|---|---|---|
| Date | The Date that the sale of a product was recorded | 26-Apr-21 |
| Retailer | The Name of the retailer that the sale of product was recorded in | Tesco |
| EAN | A unique individual article number that identifies each product | 0690251110476 |
| Description | A brief description of the product | FRAGRANCE |
| Observed ASP | The Recorded Average Selling Price a product or a group of products sold for | € 279.00 |
| Observed Units | The Recorded Number of products or a group of products that sold | 32 |

## IV. RESULTS

The aim of the study was to hypothesize if and what external KPIs might be most useful from social media data to enhance current AI and Analysis tools. Explained below is some findings on a use case for a UK based luxury department store over a two-year period from 18th May 2020 to 8th November 2021. The initial results looked at the volume of Tweets against sales recorded.

Figure 2 shows initial results, over a 6-month period, representing a spike in sales for a given retailer and the number of Tweets posted about that retailer at the same time. The lack of correlation in the data showed that peek in sales may not have been consumer driven due to little change in number of Tweets published at the same time. To gain more insights, data for sales/related Tweets over a longer period of time was required.

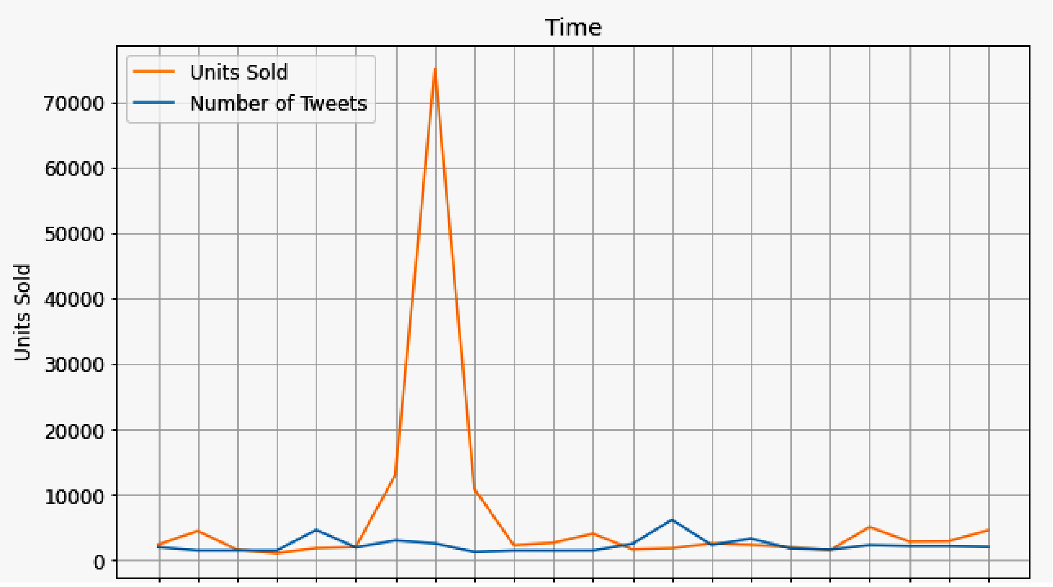

Figure 2. Units Sold & Number of Tweets – 6 Month Period

After gathering the required information, plotted in Figure 3 is a further breakdown of number of Tweets in English and number of Tweets in all other languages and the number of units sold weekly for the given retailer over a 2-year period.

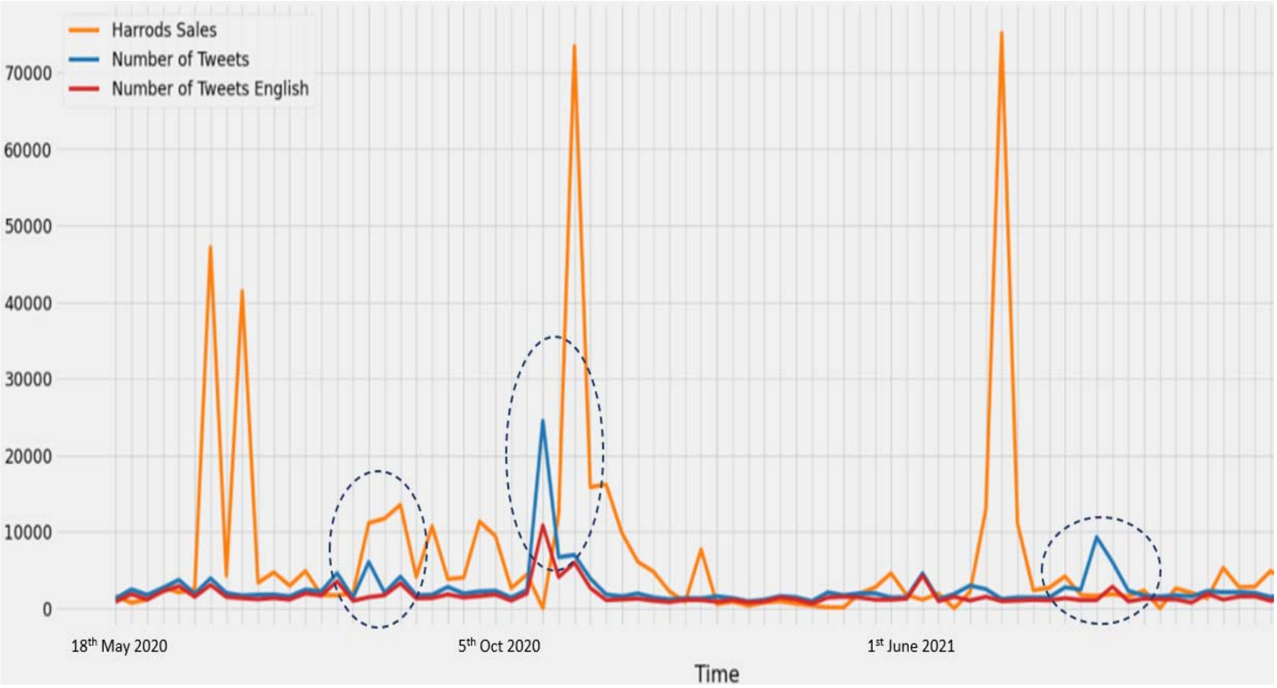

Figure 3. Units Sold & Number of Tweets – Two Year Period

Identified after analyzing the results were four main points:

- The 3 large spikes in sales are evidently abnormal and can be ignored for this use case.
- The first area of intrest highlighted in figure 4 shows an increase in Tweets in languages other than English and a rise in sales recorded and is something to be investigated further.
- The next section highlighted was an increase in Tweets in all languages including English and a rise in sales recorded and again is something to be further analysed.
- The final section to be futher looked at, highlighted is where shown to have an increase in Tweets in languages other than English but no obvious change in sales.

Figure 4 shows a plot after removing the unusual peeks of sales and looking at only observed units under 25,000. We can see now see in the data there is a much more variation and there is potential insights to be gained utilize socail media data.

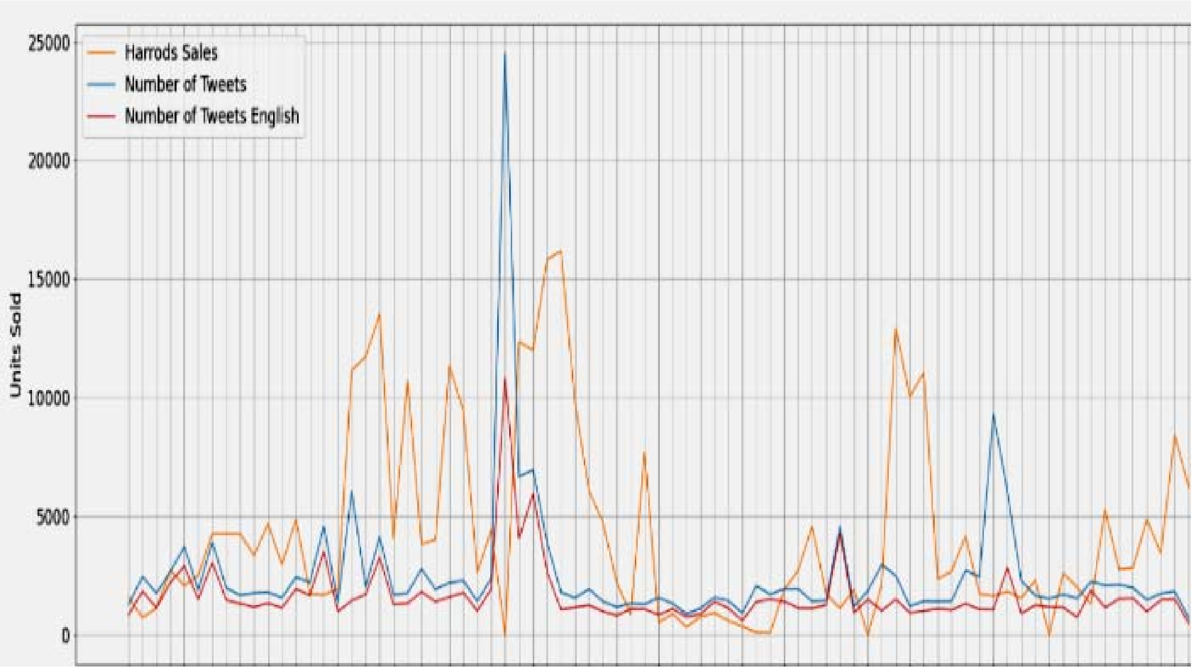

Figure 4. Abnormal peeks removed

Table IV below is a summary of statistics for the collected tweets and recorded sales.

TABLE IV.

| | |
|---|---|
| Number of overall Tweets: | 199,896 |
| Number of Tweets in English: | 127,895 |
| Number of sales recorded: | 332,344 |

## V. Future Work

After identifying in a case study over a 2-year period there is evident correlations and abnormalities in just extracting one KPI from social media data, being volume of Tweets on a given UK retailer, against the sales of that store shows there is potential for further analysis. The use of AI tools along with social media data and other external KPI's along with information from retailers could provide useful benefits in research in which this paper aimed to lay the basis for.

Future work will look at the introduction of new analysis methods to gain insights into consumer buying patterns based on contextualized user driven data and external KPIs and aiming to utilize extracted information provided by NPD along with the other KPIs such as GPS location, weather, social media data and time of year of the purchases. The selection of appropriate mechanisms for clustering, to gain insights on the data, such as KMeans, SOM and Mean-Shift will also be investigated, with the potential to train models with large sets of historical data and test on latest social media and retailer data.

## References

[1] Gerber, M.S. (2014). Predicting using Twitter and kernel dencity estimation.

[2] C. E. Lopez, M. Vasu, and C. Gallemore, “Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset,”

[3] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour,

"Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter,"

[4] T. Alshaabi, J. Minot, M. Arnold, J. L. Adams, D. R. Dewhurst, A. J. Reagan, R. Muhamad, C. M. Danforth, and P. S. Dodds, "How the world's collective attention is being paid to a pandemic: Covid-19 related 1-gram time series for 24 languages on twitter.

[5] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, "" go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19,"

[6] K.-C. Yang, C. Torres-Lugo, and F. Menczer, "Prevalence of low-credibility information on twitter during the covid-19 outbreak,"

[7] Catherine Ordun, S. P. (2020, 5 6). Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs.

[8] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A first look at covid-19 information and misinformation sharing on twitter,"

[9] Rui Xie "A Brief Analysis of Retail Customer's Consution Expierience under the Background of Artificial Intelligence"

[10] Prajogo Atmaja, Dalta Imam Maulana, Trio Adiono "AI-based Customer Behavior Analytics System using Edge Computing Dev