UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI “M. FANNO”

CORSO DI LAUREA MAGISTRALE IN
BUSINESS ADMINISTRATION

TESI DI LAUREA

ANALYSIS OF UNSTRUCTURED SOCIAL MEDIA DATA: EMPIRICAL RESEARCH ON THE WALLSTREETBETS SUBREDDIT

RELATORE:
CH.MO PROF. FABIO BUTTIGNON

LAUREANDO: EMANUELE DONÀ
MATRICOLA N. 2002783

ANNO ACCADEMICO 2023 – 2024

Dichiaro di aver preso visione del “Regolamento antiplagio” approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione ‘Riferimenti bibliografici’.

I hereby declare that I have read and understood the “Anti-plagiarism rules and regulations” approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section ‘References’.

Firma (signature) …………………………………… Emanuele Dorà

ABSTRACT

Several economic online newspaper articles between May and June 2021 blamed the sudden changes in share prices of some listed companies on the influence of social media. In this research, through text analysis, it is verified and confirmed that from the approximately eight hundred thousand comments extracted using the API wrapper PRAW of the WallStreetBets subreddit general discussions, such as "Daily Discussion Thread", "What Are Your Moves Tomorrow", "Weekend Discussion Thread", a strong interest from users emerges towards the aforementioned companies as they are highly cited. The growing increase in this interest is also accompanied by the increase in the sense of belonging to the group of redditors which could have given rise to the organization of collective actions. From the sentiment analysis on a daily basis using the dictionary-based method, it turns out that as the positivity of users' sentiment increases, the returns of the most mentioned stocks increase and vice versa.

TABLE OF CONTENTS

SECTION II - EMPIRICAL RESEARCH AND RESULTS

INTRODUCTION

Most organizations in all sectors obtain information and knowledge from predominantly structured data – typically categorized as *quantitative* data – namely in standardized formats ready for analysis. IT experts state that, globally, most of the data is unstructured – typically categorized as *qualitative* data – and, therefore, not usable as it is. Their share is between 80-90% and, among this data, 90% has been created recently while only 0.5% is used after analysing it. According to the International Data Corporation (IDC) forecast, the situation will remain so until the end of 2025 (Dialani 2020).

The massive use of the Internet generates an infinity of unstructured data, which means that it requires a standardization process to be analysed (Congruity 360 2023). IDC projections predict that by 2025, 175 zettabytes (175 trillion gigabytes) of data - much of which will be unstructured - will be generated globally (Coughlin 2018). According to Everest Group 2021, unstructured data will grow by 55-65% annually, which is triple the increase in structured data. The amount of global data created has exceeded available storage capacity since 2007. Memorizing all this data for future use is therefore very difficult but analysing it and extracting valuable information are even more difficult processes. New techniques for processing unstructured data and increasingly high-performance computers can reveal new information that would otherwise have remained hidden (Cukier 2010). Much of the unstructured data collected by an organization is textual, coming from internal and external sources such as company reports, emails, chatbot messages and information from websites and social media (Gandomi and Haider 2015). In the light of a Forbes survey, over 95% of companies claim they have problems managing and analysing unstructured data (Kulkarni 2019). As a result, they are not yet prepared to exploit the potential value of this data and thus benefit from it to support and make better forecasting choices or to identify and improve possible problems (Saggi and Jain 2018). One of the causes of this gap is the lack of expert professionals capable of analysing and interpreting unstructured data (Chen, Chiang and Storey 2012; Sagiroglu and Sinanc 2013). Even in the financial sector, it can be very useful to decode unstructured data since it can give indications on market trends, mitigate and manage risks, detect and prevent fraud (Dicuonzo et al. 2019). Part of the large flow of unstructured data can be found inside the many different social media on the Internet and they can also be exploited by external analysts.

My research is meant to analyse the unstructured data which can be found in any general discussions of particular social media, like for example Reddit, that are involved in the financial sector. Moreover, the subreddit WallStreetBets had a strong influence on the financial sector during the period from 16th May 2021 to 11th June 2021. IT tools such as APIs, specific to social platforms have been used to carry out this type of investigation, which requires data extraction to be analysed. In this research, the API wrapper PRAW, a specific API-based IT tool whose function is that of facilitating its use was employed. PRAW has to be provided with the URL of a specific webpage to extract comments from a single post: one webpage one post.

In this research, an Internet Archive tool – the Wayback Machine – was used since Reddit does not display the date of posts. Such a tool enables retrieving the stored web pages according to the precise date of publication of the posts together with its URL. Using PRAW, one could download approximately eight hundred thousand comments relating to general discussion posts, which enabled the creation of the dataset.

All research uses Natural Language Processing (NLP) to analyse unstructured data in comments. In this research, Python and one of its libraries, Natural Language Toolkit (NLTK), are used, which permits the processing and analysis of textual data. At the beginning, the first step is pre-processing the text and then, one can perform some more complex analyses, such as information extraction and sentiment analysis.

The analysis of the total dataset of comments made it possible to obtain interesting statistics on trends concerning their quantity and length, the use of emojis, and the most frequent words, many of which are tickers, that summarize the topics expressed by users. After an aggregate analysis of the comments, a deepened daily analysis made it possible to identify the way in which the most mentioned tickers are spread over the period under examination, the comparison with the relative share prices and the possible existence of an identity of group among Redditors. To assess the sentiment analysis of the comments, firstly I adopted the dictionary-based method using both the word list developed by Loughran and McDonald and the version I expanded adding new words. Then, I used the word list developed by Renault and finally, the VADER tool. The results stemming from sentiment analysis, from the number of mentions of the most cited tickers and from the group identity were compared with the returns of the most mentioned companies listed in NYSE and NASDAQ.

The thesis aims at extracting as much information as possible from the comments in the subreddit WallStreetBets through text analysis. The data so obtained could be used to verify whether some form of influence can be found with respect to the returns of the most mentioned companies by the users in the period under consideration.

The thesis is divided into two sections: a theoretical overview and another dealing with the empirical research and results. The first section comprises four chapters: the first is about social media relating to the unstructured data and the reasons for choosing the aforementioned analysed period while the second deals with IT tools to extract such data. The third chapter concerns the importance of using the Internet Archive, its Wayback Machine tool, and the dataset creation process. Finally, the fourth chapter deals with NLP, its techniques for pre-processing unstructured data, and its applications for extracting information and carrying out sentiment analysis.

The second empirical section is made up of three chapters. The fifth chapter analyses the whole of the dataset of comments at an aggregate level. The sixth chapter analyses the dataset daily in a more detailed way. The seventh chapter concerns sentiment analysis.

SECTION I - THEORETICAL OVERVIEW

CHAPTER 1 - Social media

This chapter initially addresses social media and then focuses on Reddit and WBS and the criteria adopted for choosing the period examined. Social media has become popular because it let people keep in touch with each other quickly and easily. It is continually growing thanks to the vast availability of devices, such as smartphones, tablets and PCs, which enable the access to the Internet. The way users interact with each other leads to the creation of an infinite number of different contents, most of which are in textual format, such as posts and comments (Barbier and Liu 2011). The nature of data generated in social media is large, dynamic, unstructured and noisy (spam) (Gundecha and Liu 2012). Social media encourages the immediate dissemination of news and information (Lerman and Ghosh 2010). Traditional media, such as television, newspapers and radio, provide people with information that is mainly unidirectional since it comes from one single source and directly reaches many consumers. On the contrary, social media has revolutionized the usual ways of obtaining information: anyone can publish various types of content that can reach large masses of people quickly, thus apparently democratizing information and influencing the individuals' thinking and behaviour (Barbier and Liu 2011).

This research takes into consideration Reddit, a social media that has gained increasing popularity in recent years and has significant and concrete implications for the financial sector. Following these events, there has also been a growing interest in the academic field (Nobanee and Ellili 2023).

1.1 - Reddit

Reddit is a social news, web content discussion and aggregation site founded on June 23rd in 2005, in Medford (Massachusetts), stemming from the idea of Steve Huffman and Alexis Ohanian, two young graduates. The name Reddit comes from the union of two English terms - *read/edit and read it* - whose primary function is that of reading topics and enriching the discussion through posts and comments (Fastweb 2022). Reddit's original motto is “The Front Page of the Internet” (Ohanian 2013), as it reports the most important news and contents of the day determined by the users of the site just like a newspaper would do. Reddit was first written in the Common LISP programming language but was later rewritten in Python because it offered more libraries and more development flexibility. He was later replaced by Pylons (Ionos 2019). Experts indicate that Reddit's popularity is due to the fact that freedom of speech is preserved, and users can browse and publish their content anonymously. In fact, unlike many other social networks, Reddit can only be used with a username and an email address. Other reasons that make it so popular and used daily by millions of users may be the presence of very specific communities and good levels of moderation (Ionos 2019). Freely registered Reddit members, called redditors, can post content such as images, links and text messages, which people can either upvote or downvote, comment on and share. Even those not registered, called lurkers, can access Reddit, but the functionality is limited to reading only.

In the same way as Redditors, lurkers can be regarded as silent participants although not intervening in discussions (Nonnecke and Preece 2000). The platform is organized into subreddits, called communities, comprising user-created boards covering many topics. Redditors talk about the different issues that the specific subreddit deals with. Users can start their own subreddits on various topics: if other users find them interesting, they can subscribe to them and thus, receive updates (Fastweb 2022). It is considered inactive if a subreddit has not had any discussions for a long time. In fact, Reddit considers a subreddit to be active when it receives at least five comments per day (Lin 2023). The number of subreddits has grown steadily since Reddit was founded. There were approximately 2.8 million subreddits in 2021 (Metrics for reddit 2023) and approximately 3.1 million in 2022 (Dean 2023). The official Reddit website (Reddit 2024) reports that in October 2023 there were more than 100 thousand active communities, more than 16 billion posts and comments and more than 70 million daily active users, while 52 million in 2021 (Dixon 2023a). Reddit is one of the most visited websites online. According to Similarweb (2023), total visits to Reddit were around 2.1 billion in December 2023 while in 2021 they were 1.68 billion. Reddit is ranked as the 16th most visited website in the world, the 9th in the United States, and the 4th among similar sites in its category. Again, according to Similarweb (2023), in December 2023 the majority of Redditors reside in the United States with 48.51%, followed by the United Kingdom with 7.15%, Canada with 7.01%, Australia with 4.36%, Germany with 3.06% and other countries with 29.91%. So, most users reside in predominantly English-speaking states. According to Statista (Dixon 2023b), in 2022, the 63.8% of global Reddit users are male while the 36.2% are female. So, it can be said that Reddit is generally more prevalent among men. Furthermore, in 2021 in the United States the 36% of Redditors belonged to the 18-29 age group, the 22% to the 30-49 age group, the 10% to the 50-64 age group and the remaining 3% to the over 65s (Dixon 2022c). Instead, in the United Kingdom the most significant number of Redditors is aged between 15 and 35 with 47%, between 36 and 55 with 17%, while only 1% is over 56 (Dixon 2022b). Moreover, according to Statista (Dixon 2022f), it appears that in 2021 in the United States most adult Reddit users live in urban and suburban areas and have attended college (Dixon 2022e). Furthermore, most fall into the upper-middle income bracket (Dixon 2022d). Finally, data from Statista

(Dixon 2022a) shows that in the third quarter of 2020 in the United States, the 52% of Reddit users accessed the platform daily, 82% accessed it weekly, the 95% accessed it monthly. On the Reddit platform, in some subreddits such as WallStreetBets, topics related to economics and finance are addressed, in which there are discussions on issues relating to investments and cryptocurrencies.

1.2 - WallStreetBets

Thanks to the WallStreetBets subreddit, hereafter WSB, there has been a further increase in Reddit's popularity after the GameStop case. In fact, in 2021 the three most popular posts on Reddit came from the r/wallstreetbets subreddit (Dailey 2021). WSB is a subreddit where 15 million potential investors are currently registered and discuss trading strategies, investments and their profits and losses. Globally Compared to many subreddits WSB's strength lies in its large number of subscribers, in fact it is in the top 1% of the ranking based on size. The language used in the conversations of this subreddit, which reinforces the identity of belonging (Bernstein et al. 2011; Lucchini et al. 2021), is characterized by an irreverent and informal style, full of emojis (Emoji combos 2024; Emoji party 2024; Emojipedia 2024; Symbl 2024) (Appendix A), slang, acronyms, and own terms (Reddit 2017; Reddit 2021; Business English 2024) (Appendix B). WSB is a community without any single leader (Anand and Pathak 2021) but once the discussion turns out to be persuasive, heterogeneous opinions become homogeneous and the community comes to an agreement forming polarized opinions that trigger possible collective actions (Lucchini et al. 2021). In recent years, WSB has given rise to several online mass collective coordination, capable of having a significant impact on the financial markets. From these collective actions the term ‘meme stock’ was born. This term refers to the shares of a particular company on which most of the discussions and interests of retail investors in social media converge (Aloosh, Ouzan and Shahzad 2022). The collective action of these subjects determines a rapid and significant increase in the price of security and in trading volumes. Potential WSB investors, aware that by uniting in collective actions enable them to influence the share price, channelled their activity into the mass purchase of shares of mainly troubled companies. In this way, they created significant problems for short sellers (hedge funds) who bet on the failure of these companies. The mass buying of those shares increased their prices due to excess demand and lack of availability. This forced short sellers to buy shares to limit their losses, thus driving prices even higher (Allen et al. 2021; Lucchini et al. 2021). There are several cases in which some companies have been subjected to meme stocks following this modus operandi and the best known are the following: GameStop, AMC Entertainment, BlackBerry, and Nokia (Lyócsa, Baumöhl and Výrost 2022).

1.3 - Choice of the period

As for the period to analyse, researching into economic online newspaper articles provided various information relating to social media and their possible influence on financial markets. In particular, the choice fell on the examination of articles that highlighted cases of companies that had had sudden increases and rapid drops in share prices, and that attributed the responsibility for these events mainly to the influence of social media. The question arose as to whether this price trend could really be traced back to social media focusing on the financial sector, particularly WSB. Once the different news reports had been consulted, the investigation had to be conducted on extended potential periods since not only a few significant isolated cases had occurred, but also others that had received less media coverage (La Monica 2021; Miao and Stevens 2021; White 2021). The choice of the period - from the 16th May2021 to the11th June 2021 - happened once the cases of interest involving different companies had been detected. In this way, several days preceding the sudden increases in share prices were included and a few days later their rapid collapses. The analysis of the comments published in WSB in that specific period can provide valuable information on the discussions, sentiment and strategies implemented by the community, thus revealing how much and how it influenced the financial market. In order to investigate the discussions that took place in social media and the phenomena connected to them, specific tools are needed that allow the extraction of the relevant data. The Application Programming Interface allows you to carry out this task.

CHAPTER 2 - API and API wrapper

Downloading comments from social networks involves interacting with the platform's own API (acronym for Application Programming Interface), or the use of specialized tools that allow access to public data, such as API Wrappers.

An API is a set of communication protocols, routines (sequence of instructions) and tools that allow devices and software applications or different applications to communicate and interact with each other. An API can be built using various programming languages. APIs then define the methods and data formats that applications have to use to request and exchange information. They enable the interaction of different software systems allowing them to work together and share data quickly, efficiently and securely. This interaction can involve capturing data, sending data, or performing specific actions. Instead of creating internal software a company can use an API to make its software programs or data interact with that of another company. This entails a significant saving of time and money. APIs are commonly used in web development to enable communication between web servers and client applications but can be found in various software contexts. (Rapidapi 2023) An example of an API between web server and client is the payment of a product purchased on a website via mobile phone or computer through one of the digital payment systems. The entire process is carried out through APIs, which exchange data without the user having visibility at the interface level.

This is the process that happens between client and server:

1. The client (system or application) sends the request to the API server - called "API call"- which specifies the operation to be carried out;
2. The API server receives the request and, processes it using the application;
3. The API server returns a response to the client formatted according to the API specifications, with which it communicates the requested data;
4. After receiving the response from the API, the client processes the data according to the programmer's method, for example, showing it to the user (IBM 2023).

APIs can be:

- private (or even internal), when they are used only within a company to connect systems and data, and remain hidden from external users;
- of partners, when they are accessible only to trusted partners external to the company to facilitate connections with other companies;

- public (or even open or external), when anyone can use it with any authorizations and/or costs, thus allowing third-party developers to create applications that interact with company APIs. The more an API is made public the more developers are inclined to develop applications centred on this API;
- composite, when different APIs are composed together in complex systems. Developers can thus obtain data from multiple sources through a single API (Lutkevich 2022).

APIs offer numerous benefits. First, their functionality since they facilitate the development of the application with which information is exchanged between different systems, adding new services. Then their modularity, which allows a complex system to be divided into independent modules, thus facilitating the management, maintenance and updating of individual parts without having to start from scratch. Another benefit is the reusability of individual software components which allows us to reduce the work required for their design, thus encouraging the development of new features. Interoperability, then, allows the interaction among various systems to be standardized, facilitating the integration between different technologies and services. Finally, the use of APIs facilitates accessibility to services and guarantees the security of services and data (Devinterface 2023). Specifically, the Reddit API allows developers to access and interact with Reddit data, including reading and submitting posts, retrieving comments and more (Reddit 2023).

In the IT field, an API wrapper is a package of code that allows different programs to communicate with each other, simplifying the use of a specific API. API wrappers handle the details of making requests to an API, parsing responses and other low-level interactions. They can be low-level when they enable a direct interface with a system, or they can be high-level when they provide the developer with an easier way to use interface. API wrappers are valuable tools for developers because they simplify the interaction between services and APIs and facilitate integration between different technologies by improving their compatibility. Therefore, they let the user save time and effort while at the same time promoting consistency and good practice. Among the many crucial aspects of API wrappers, there is the fact that they often handle the authentication process making it easier for developers to obtain and use the necessary credentials. They also extract the underlying details of HTTP request execution and response analysis in a more user-friendly, high-level interface. API wrappers can include API error and exception handling mechanisms to simplify the process. They often enable common tasks and operations by reducing the amount of code required to perform specific actions making it easier for developers to integrate API functionality into their applications. Finally, API wrappers aim at providing a consistent and predictable interface and the best of them come with complete documentation and examples to guide developers on how to use the library effectively and efficiently (Got API 2023).

To sum up, while an API is a set of rules and protocols for the interaction between devices and software components, an API wrapper is a tool or library obtained using one of the possible programming languages that simplifies the use of a specific API making it more accessible and convenient for developers.

2.1 - Choosing an API wrapper (PRAW)

The choice of a specific API wrapper depends on its ease of use, on the continuity in development, on community support, on its functionality, and on personal preferences. It also depends on the requirements of one’s own project, familiarity with the library and on the presence of specific features that any library offers compared to another. It is also essential to review the documentation and features of different wrappers to determine which one best suit one’s needs (Manaw 2023).

PRAW - acronym for "Python Reddit API Wrapper" - is a package (series of modules) developed using the Python language, which enables an easy access to the Reddit API following all its rules (PRAW 2023a). PRAW is widely used for several reasons, firstly as it provides a Python interface that is widely used for its easiness, which allows developers to interact with the Reddit API without having to deal with low-level details. The syntax is designed to be user-friendly, thus making it accessible to beginners and experienced developers. Secondly, PRAW simplifies the use of the Reddit API, handling tasks such as authentication, managing HTTP requests and parsing JSON responses. This allows developers to focus on their specific task creating applications or scripts that leverage Reddit data rather than dealing with the complex details of the API. Moreover, PRAW offers a range of features such as reading posts, posting comments, managing user authentication and more. Additionally, PRAW is extensible allowing developers to customize and extend its functionality. It also includes built-in support for managing Reddit API rate limits helping developers avoid issues related to excessive requests. The developer community is active and supportive, and it is invaluable to get help and solve the different problems that may arise. The library is well documented thanks to comprehensive guides and examples, which let users understand and implement Reddit API interactions. Being actively maintained and receiving regular updates, PRAW ensures that the library keeps compatible with changes to the Reddit API and keeps improving over time. Thanks to regular updates users can benefit from bug fixes and new features. Finally, PRAW is an open-source project, that is the source code is freely available, and can be modified, expanded, and redistributed without any restrictions. This openness fosters transparency and collaboration within the developer community (Eliahu 2023).

The Reddit API and in turn PRAW allow obtaining all comments from submissions which will then be used to perform text analysis. The submission can be specified through the entire URL, acronym for Uniform Resource Locator (for example https://www.reddit.com/r/funny/comments/3g1jfi/buttons/), or a part of it containing the ID (for instance 3g1jfi which follows comments/) and which represent a specific post in a unique way (PRAW 2023b).

2.2 - Limitations of the analysis by time period and keywords

Through the use of some specific API wrappers, all the texts on a particular site could be identified and downloaded by filtering the posts according to a precise period of time and certain keywords, such as a company ticker (Pagolu et al. 2016). However, since not all API wrappers are one and the same thing some of them may not have this specific functionality. These are the most critical limitations that some API wrappers may meet whenever they filter posts based on particular time periods and keywords (Cruz, Kinyua and Mutigwe 2023). Using only a few keywords, without considering different terms or expressions that can refer to the same topic, such as synonyms, abbreviations, or alternative expressions, may result in loss of relevant content. Furthermore, false positives and information noise could also be detected in the data set based on specific keywords. In the first case it is data that, although responding to the request, does not refer to the expected topic while in the second, instead, refers to data that, although dealing with what is wanted, are not relevant (Agenda Digitale 2018). Language proficiency is another aspect that can cause limitations, also in reference to the period. As a matter of fact, we cannot always have knowledge of all the words that are used in different contexts or phenomena and so, consequently, the risk is to leave out words that we do not know but that are relevant and conduct an incomplete search, with partial or incorrect results. Finally, language is dynamic by its nature and changes over time: discussions may use variations, slang or new terms that are not covered by the selected keywords. Keeping up with changing language trends requires regular updates to the keyword list.

Given the limitations of the approach just mentioned, in the data extraction phase of this research, it was decided not to apply period and keyword filters but only to extract general discussion posts.

2.3 - Importance of daily discussion threads pinned by moderators in Reddit

In social media, a thread is a discussion that takes place on a well-defined topic. A first user, generally the moderator, establishes the subject – called topic – of the discussion; then, the users' interventions on the topic or the answers given by other users follow in chronological order. The set of topics and interventions is called a thread (Hootsuite, 2020).

In this research we chose to analyse the daily discussion threads of the WSB subreddit rather than posts containing one or more keywords because analysing daily discussion threads allows to enter into the most visible and quickly accessible conversations within the same community. As for specific particular research objectives, it can be very useful more to analyse a daily discussion thread than posts containing one or more keywords, providing a snapshot of collective thinking on a specific aspect or topic. These threads often follow a definite structure determined by the community moderators following strict rules: this makes the analysis more manageable and rigorous and allows insights into the most discussed or trending topics. In addition, the investigation of the entire thread allows you to directly acquire a broad view of the sentiment, discussions and opinions of the community on that particular topic or theme. The regular review of daily threads, in which discussions on the same or similar topics over different periods are compared at different times, can help you understand how the overall sentiment is changing and identify emerging trends. Finally, the analysis of daily threads can help to understand the dynamics that develop continuously within the community; it let us observe and understand how users interact with each other, which topics generate the most outstanding involvement and how feelings can evolve over time.

Not all Reddit pages use a daily discussion thread, which depends on the policies, nature, focus, and specific preferences of the moderators and the community members. Daily discussion threads are commonly used in certain types of subreddits, such as those dedicated to particular topics, hobbies, or activities where regular, informal conversations between community members are encouraged. For example, subreddits related to financial markets, fitness, gaming, and other interests often have daily discussion threads where users can share thoughts, ask questions, and engage in free conversation. Many WSB pages also contain threads such as '*Daily Discussion Thread'*, '*What Are Your Moves Tomorrow*', '*Weekend Discussion Thread*' characterized by a yellow or blue label, with which Reddit indicates that moderators of a particular subreddit have chosen to highlight and pin a certain post or comment at the top of the main page (Mancini et al. 2021, Boylston et al. 2021). As a result, displaying them first they make easily accessible some information deemed essential, such as announcements, discussions, frequently asked questions, ongoing events, community guidelines or other content so that visitors to the subreddit will see the pinned post first, although others will be posted later on the same page. The number of pinned posts that can be displayed at the top of a subreddit is usually limited to one or two, depending on the settings that have been given.

For a better understanding of what was happening and to have a clear overview of the WSB subreddit in the period under review the Internet Archive – an important resource – was used.

CHAPTER 3 - Internet Archive and dataset

Internet Archive is a non-profit digital library founded in 1996 that aims at providing "universal access to all knowledge" (Archive 2014). It collects millions of books, films, software, music, websites and more giving anyone the opportunity to access content for free. The original mission of the Internet Archive, as the name suggests, is to store the whole of digital contents on the Internet. Unlike a newspaper or a printed book, these contents are intangible and, therefore, subject to modifications and even to permanent deletion at any time. This is why there is often no memory left of them, not even of aesthetic nature. In fact, the average life of a web page, prior to modification or deletion, varies from forty-four to one hundred days (Ashenfelder 2011). The Internet Archive preserves digital documents for future generations and provides historical documentation of the web, which is constantly evolving by its own nature. Internet Archive gives the possibility to search, through an integrated tool introduced in 2001 - the Wayback Machine - a history which today is made up of over 860 billion web pages on the Internet. Storing on the Internet Archive enables the consultation of web pages that either have changed over time or are no longer reachable through regular search engines because they have been closed (Archive 2022a.). Harvard University conducted a study on over 550,000 New York Times articles published online from 1996 to 2019 and discovered that on average 25% of the links contained in them are no longer working, meaning that the pages to them connected have been deleted or modified. Broken links (Rouse 2015a) went from the 6% of connections in 2018 to 72% in 1998 (Clark, 2021). Another study carried out in 1995 on 360 links on the Internet discovered that 20 years later, only 1.2% of them were still working (Rais, 2022).

3.1 - The subjects who contribute to the creation of the archive

The web pages that Wayback Machine can visualize are saved and stored by means of a combination of automated scans, of user submissions, and of with website owners collaborations. Users actively contribute to the archive by submitting specific URLs via the "Save Page Now" feature on the Wayback Machine website. This allows anyone to store pages they consider necessary or exciting. Additionally, website owners can collaborate with Internet Archive by providing archives of their sites (Archive 2022b). The automated programs managed by Internet Archive that analyse the contents of the network, called web crawlers, keep browsing the web storing snapshots of pages and indexing their contents.

Web crawlers – also known as web spiders, web robots or simply bots – are automated programs or scripts designed to navigate the World Wide Web by systematically and continuously collecting information from websites. They analyse keywords, the various contents present, the internal and external links and index them, which means that they catalogue the collected information by entering it into the search engine database. Therefore, these crawlers play a fundamental role: they allow search engines to collect data on web pages, to constantly update their databases and to make information easily accessible to users through search results (Rouse 2017). Web crawlers are widely used by search engines such as Google, Bing and Yahoo to index the vast amount of information catalogued and available on the web (Rouse 2015b; TechTerms 2008).

The web crawler process generally occurs in three phases. The first one consists in scanning and archiving web pages. Using web crawlers similar to those used by search engines, Internet Archive visits and scans websites on the Internet, capturing their entire content, including HTML tags, images, style sheets, scripts, links, and other resources associated with the page. Crawlers also follow links from one page to another and finally, they store what has been acquired (Lyden 2017). Then, the second step is the indexing, which makes the acquired content searchable via Wayback Machine. By entering a URL, users can view archived versions of the corresponding web page (Morris, 2023). In the end, each version of a web page captured by the Internet Archive is given a timestamp, which allows users to log in and view the page just as it appeared (Graham 2017).

The frequency with which web pages are acquired by Internet Archive web crawlers can vary for several reasons, one of them being the popularity and importance of the site. High-traffic pages and websites the historical, cultural, or academic value of which is significant can receive great attention from Internet Archive web crawlers. In addition, since the purpose of the Internet Archive is to provide a historical record of web content, pages that undergo continuous changes or close updates are likely to be captured more frequently than others so, they can have multiple archived versions, even on the same day. The Internet Archive web crawlers also follow the links among the various pages: whenever a page has more incoming links or is part of a well-connected network of pages it can be crawled and acquired more frequently. Another aspect concerns the ability of website owners to control whether their sites are crawled and stored by the Internet Archive through the use of the "robots.txt" file. If a website owner requests exclusion, the Internet Archive comply with this directive and may capture the page less frequently or not at all. It should be noted that some pages could encounter technical difficulties for web crawlers affecting the frequency of successful acquisitions. For example, it can be very challenging to accurately store pages with complex JavaScript interactions or other non-standard elements. Finally, since the Internet Archive has limited resources the frequency of storing may be affected by the capacity for the organisation to run a scan and archive web content. Therefore, higher or more accessible priority pages could be scanned more frequently.

3.2 - Relevance and use of Wayback Machine

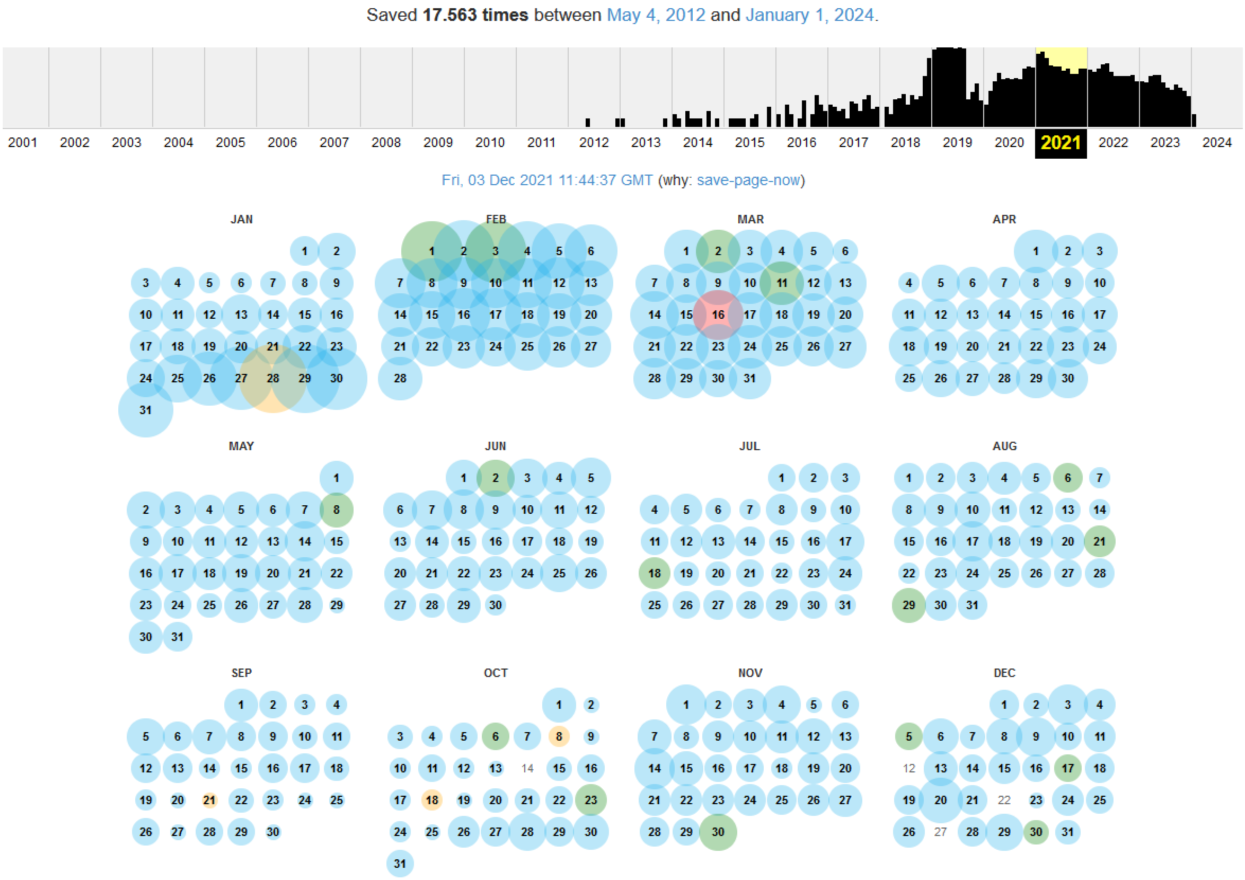
The following paragraph deals with both some specific features of the Wayback Machine and the web pages stored by the Internet Archive that make this tool particularly useful together with an insight into the way it can be used.

The analysis of the various snapshots of the same web page and their comparison can be useful to identify changes and updates to the contents, and in particular to track news websites, blogs or other platforms where information is updated regularly and to follow the evolution over time. Examining archived web pages in relation to specific dates or events, such as major news events, technological advances, or cultural changes, can offer historical perspective on how the web responded to these particular events. Moreover, these archived web pages can be analysed to study cultural and social trends emerging from online content. Research might address changes in the use of a particular language, the emergence of new topics or common interests and changes in public sentiment.

Internet Archive captures content as diverse as news articles, academic documents, blogs, multimedia (video, photographs, music, text), and more. The analysis of all this content through the Wayback Machine can provide interesting data on the kind of information that people create and share on the web, which in turn can be processed. The Wayback Machine can also be helpful to investigate the disappearance of entire websites, parts of them or specific contents from the active web, the reasons for which can be various going from a new design of the site to the change of the domain name or its expiration; from the deliberate removal of contents due to incorrect information to erroneous assessments of particular events; from the lack of authorizations to controversies of a judicial or other nature.

The quickest way to search in Wayback Machine is to type a specific URL and, in this particular research the choice fell on https://www.reddit.com/r/wallstreetbets/. The image below (Figure 3.1) shows what appears on the screen soon after typing the URL and pressing ‘enter’. On top of the page, there is a horizontal strip that represents the time frame in which the scans present in the Internet Archive were stored divided into years. Within each year, the histogram shows the quantity of scans carried out divided into months: the higher the black line the more significant their number. These scans cover a period of about 13 years: from 2012 – the year of foundation of the WSB subreddit was founded, until today. By clicking within a specific year (2021 in this case), at the bottom of the page a whole calendar in the traditional calendar format appears, displaying the different scans that have taken place, highlighted using specific colours (Archive, 2023).

Figure 3.1: Scan of the URL https://www.reddit.com/r/wallstreetbets/ in Wayback Machine.



Saved **17.563 times** between May 4, 2012 and January 1, 2024.

Fri, 03 Dec 2021 11:44:37 GMT (why: save-page-now)

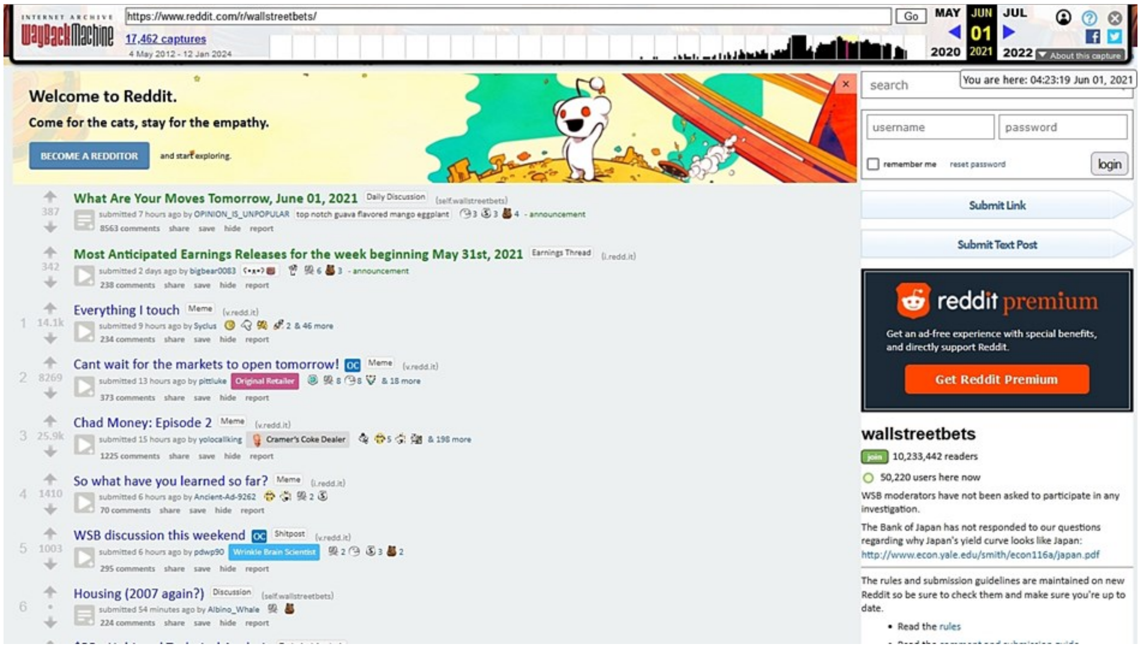
Source: archive.org

The circled days correspond to the days when the site was crawled: the larger the circle the greater the number of scans that have been performed on that day. The circles can be blue, green, orange or red: blue means that the scans occurred without any problems; green means that in that particular day that precise acquisition was a redirection; orange means that the URL was not found and finally, red means that, while retrieving the page, the server has made some errors. The majority of the circles are highlighted in light blue (Graham 2016). By positioning the mouse on one of the days when one or more scans have taken place, a window will open displaying their precise number of scans that occurred in that day together with the time when they have been performed. By clicking on the desired time, the respective snapshot will be displayed.

3.2.1 - From Internet Archive to WallStreetBets as it was and as it is now

By placing the mouse on a specific day of the calendar a window will open displaying the list of all the scans and of the time when they have taken place. For example, by opening the first of the WSB subreddit scans on 1st June 2021, this is what appears on the screen (Figure 3.2):

Figure 3.2: Example of screen after selecting a specific scan in Wayback Machine.



Welcome to Reddit.

Come for the cats, stay for the empathy.

and start exploring.

387 What Are Your Moves Tomorrow, June 01, 2021 Daily Discussion (self.wallstreetbets)
submitted 7 hours ago by OPINION_IS_UNPOPULAR top notch guava flavored mango eggplant 3 3 4 - announcement
8563 comments share save hide report

342 Most Anticipated Earnings Releases for the week beginning May 31st, 2021 Earnings Thread (i.redd.it)
submitted 2 days ago by bigbear0083 6 3 - announcement
238 comments share save hide report

1 14.1k Everything I touch Meme (v.redd.it)
submitted 9 hours ago by Syclus 2 & 46 more
234 comments share save hide report

2 8269 Cant wait for the markets to open tomorrow! OC Meme (v.redd.it)
submitted 13 hours ago by pittluke Original Retailer 8 8 & 18 more
373 comments share save hide report

3 25.9k Chad Money: Episode 2 Meme (v.redd.it)
submitted 15 hours ago by yolocallking Cramer's Coke Dealer 5 & 198 more
1225 comments share save hide report

4 1410 So what have you learned so far? Meme (i.redd.it)
submitted 6 hours ago by Ancient-Ad-9262 2
70 comments share save hide report

5 1003 WSB discussion this weekend OC Shitpost (v.redd.it)
submitted 6 hours ago by pdwp90 Wrinkle Brain Scientist 2 3 2
295 comments share save hide report

6 • Housing (2007 again?) Discussion (self.wallstreetbets)
submitted 54 minutes ago by Albino_Whale
224 comments share save hide report

reddit premium

Get an ad-free experience with special benefits, and directly support Reddit.

wallstreetbets

10,233,442 readers

○ 50,220 users here now

WSB moderators have not been asked to participate in any investigation.

The Bank of Japan has not responded to our questions regarding why Japan's yield curve looks like Japan: http://www.econ.yale.edu/smith/econ116a/japan.pdf

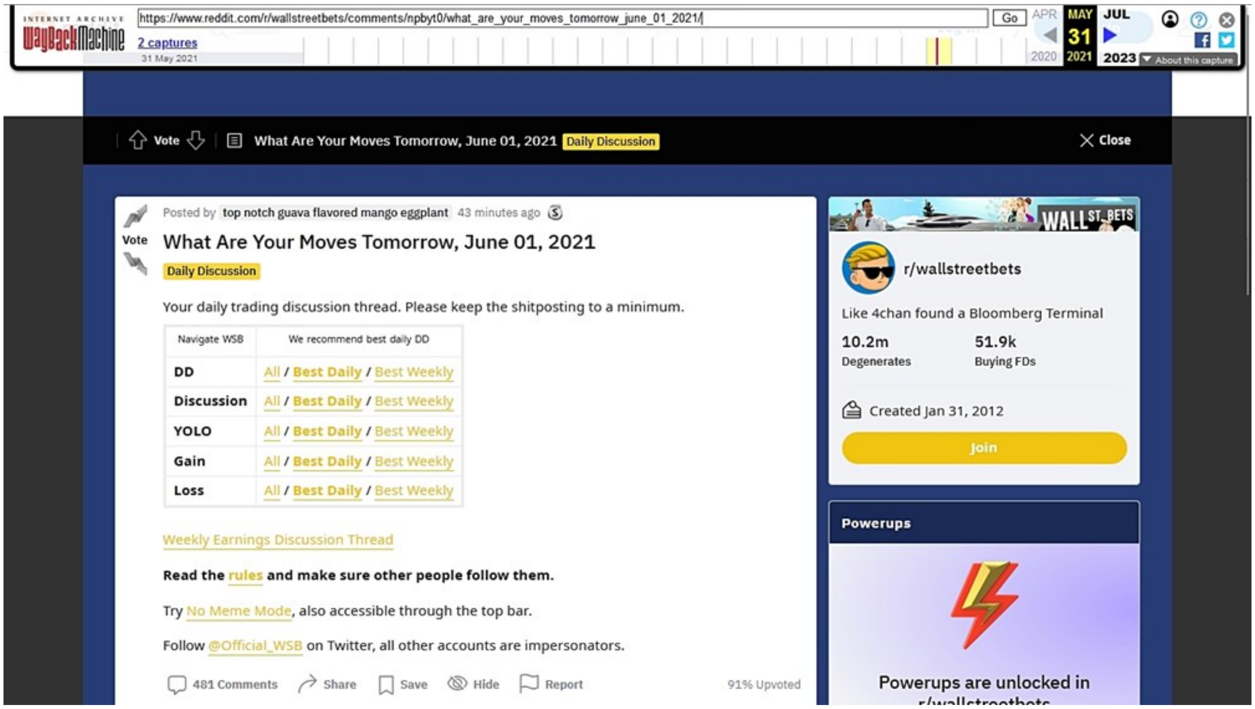
The rules and submission guidelines are maintained on new Reddit so be sure to check them and make sure you're up to date.

- Read the rules

Source: archive.org

A click on one of the headlines in the central part of the page, for example the first, (Figure 3.2), will open the link and the relevant page will be displayed (Figure 3.3). At the top of the page, the URL of the acquired page appears, which is important to visualize the current web page.

Figure 3.3: Example of screen after selecting a specific title in a scan in Wayback Machine.



https://www.reddit.com/r/wallstreetbets/comments/npbyt0/what_are_your_moves_tomorrow_june_01_2021/

What Are Your Moves Tomorrow, June 01, 2021 Daily Discussion

Posted by top notch guava flavored mango eggplant 43 minutes ago

What Are Your Moves Tomorrow, June 01, 2021

Daily Discussion

Your daily trading discussion thread. Please keep the shitposting to a minimum.

Navigate WSB	We recommend best daily DD
DD	All / Best Daily / Best Weekly
Discussion	All / Best Daily / Best Weekly
YOLO	All / Best Daily / Best Weekly
Gain	All / Best Daily / Best Weekly
Loss	All / Best Daily / Best Weekly

Weekly Earnings Discussion Thread

Read the rules and make sure other people follow them.

Try No Meme Mode, also accessible through the top bar.

Follow @Official_WSB on Twitter, all other accounts are impersonators.

481 Comments Share Save Hide Report 91% Upvoted

WALL ST BETS

r/wallstreetbets

Like 4chan found a Bloomberg Terminal

10.2m Degenerates

51.9k Buying FDs

Created Jan 31, 2012

Join

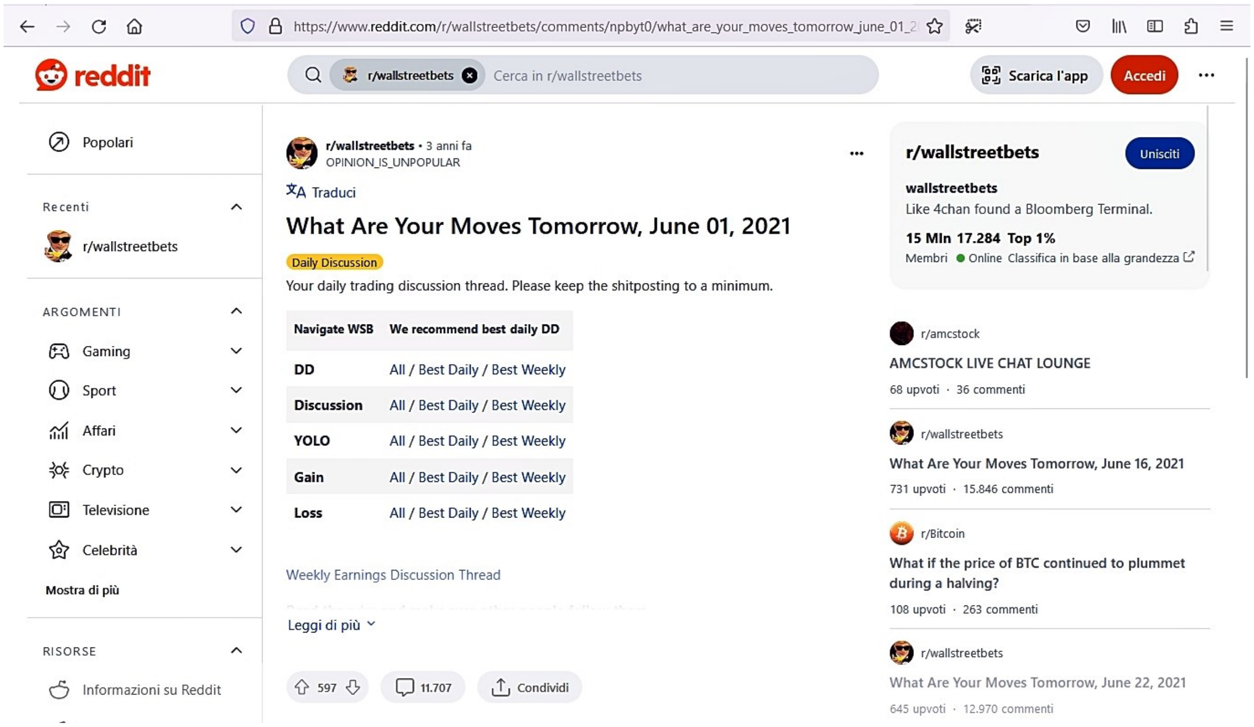
Powerups

Powerups are unlocked in r/wallstreetbets

Source: archive.org

Copying and pasting the above-mentioned URL into the search engine bar and pressing Enter will open the page as it appears today (Figure 3.4). This URL is essential in order to extract the comments of the specific post, as it is specifically requested by Reddit's API

Figure 3.4: Example of the screen of how the web page looks today.



r/wallstreetbets • 3 anni fa
OPINION_IS_UNPOPULAR

What Are Your Moves Tomorrow, June 01, 2021

Daily Discussion

Your daily trading discussion thread. Please keep the shitposting to a minimum.

Navigate WSB	We recommend best daily DD
DD	All / Best Daily / Best Weekly
Discussion	All / Best Daily / Best Weekly
YOLO	All / Best Daily / Best Weekly
Gain	All / Best Daily / Best Weekly
Loss	All / Best Daily / Best Weekly

Weekly Earnings Discussion Thread

r/wallstreetbets
wallstreetbets
Like 4chan found a Bloomberg Terminal.
15 Mln Membri · 17.284 Online · Top 1% Classifica in base alla grandezza

r/amcstock
AMCSTOCK LIVE CHAT LOUNGE
68 upvoti · 36 commenti

r/wallstreetbets
What Are Your Moves Tomorrow, June 16, 2021
731 upvoti · 15.846 commenti

r/Bitcoin
What if the price of BTC continued to plummet during a halving?
108 upvoti · 263 commenti

r/wallstreetbets
What Are Your Moves Tomorrow, June 22, 2021
645 upvoti · 12.970 commenti

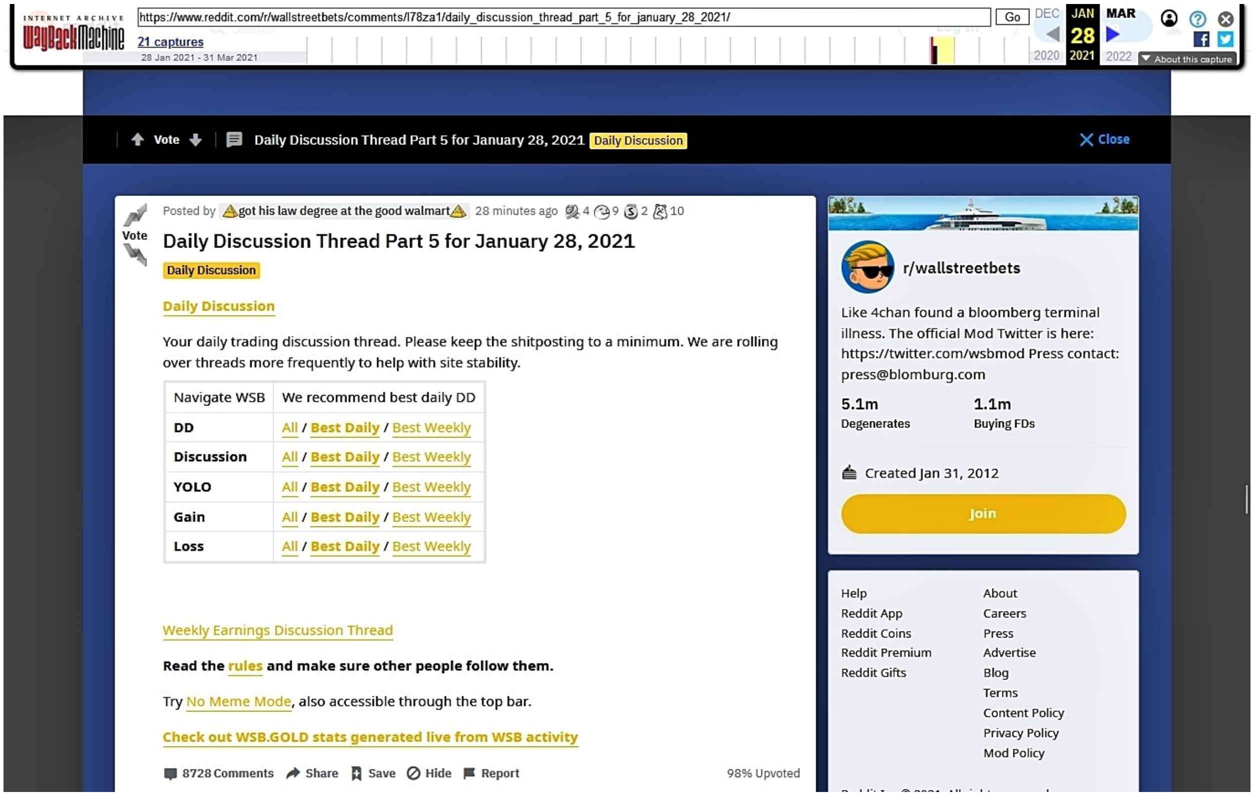
Source: archive.org

Reddit does not give the possibility to detect the exact date on which a particular post was published, nor does it allow searching for posts by filtering them by date. Thus, if an analysis is made concerning a post published long ago, Reddit will only show how many years earlier it was published, without indicating either the month or the day. To overcome this obstacle, simply put the operator 'inurl:' in front of the URL of the page in question to detect the date (Fur 2020; Titinnanzi 2018). If this fails, using the method just described, the publication date can be obtained by analysing the HTML code of the page and searching for it via the 'created-timestamp' entry (TechTerms 2021).

3.2.2 - Recovering deleted posts from Internet Archive

The Wayback Machine makes it possible to retrieve the post relating to a discussion that, although it had obtained several comments, was subsequently deleted by the author himself or by the moderators, and is now no longer visible. Taking, for example, the discussion that took place on 28 January 2021 in WSB in the Internet Archive (Figure 3.5), it can be seen that, as written at the bottom of the page, the post after 28 minutes had already obtained 8728 comments.

Figure 3.5: Example of a thread in Wayback Machine.



https://www.reddit.com/r/wallstreetbets/comments/l78za1/daily_discussion_thread_part_5_for_january_28_2021/

21 captures

28 Jan 2021 - 31 Mar 2021

Daily Discussion Thread Part 5 for January 28, 2021 Daily Discussion

Posted by ⚠got his law degree at the good walmart⚠ 28 minutes ago 4 9 2 10

Daily Discussion Thread Part 5 for January 28, 2021

Daily Discussion

Daily Discussion

Your daily trading discussion thread. Please keep the shitposting to a minimum. We are rolling over threads more frequently to help with site stability.

Navigate WSB	We recommend best daily DD
DD	All / Best Daily / Best Weekly
Discussion	All / Best Daily / Best Weekly
YOLO	All / Best Daily / Best Weekly
Gain	All / Best Daily / Best Weekly
Loss	All / Best Daily / Best Weekly

Weekly Earnings Discussion Thread

Read the rules and make sure other people follow them.

Try No Meme Mode, also accessible through the top bar.

Check out WSB.GOLD stats generated live from WSB activity

8728 Comments · Share · Save · Hide · Report · 98% Upvoted

r/wallstreetbets

Like 4chan found a bloomberg terminal illness. The official Mod Twitter is here: https://twitter.com/wsbmod Press contact: press@blomburg.com

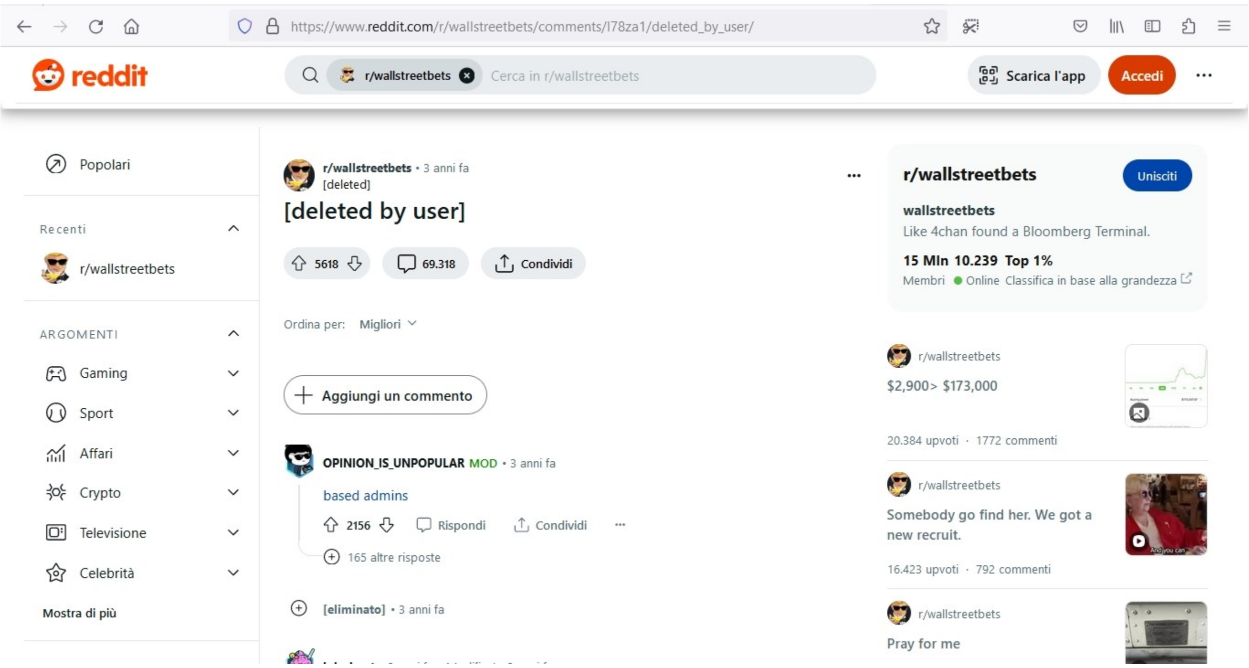
5.1m Degenerates · 1.1m Buying FDs

Created Jan 31, 2012

Source: archive.org

If you go to the address at the top of the page (Figure 3.5), you can see on the page below (Figure 3.6) that both the post and the author have been deleted, and that the comments have reached the remarkable number of 69,318. In this way, one can understand the content of the post and download comments relevant to one's research.

Figure 3.6: Example of the screen of how the web page looks today without author and topic.



[deleted by user]

based admins

Like 4chan found a Bloomberg Terminal.

$2,900> $173,000

Somebody go find her. We got a new recruit.

Pray for me

Source: archive.org

3.3 - Creation of the dataset

Once all URLs of the general discussion posts from the period under review were obtained, the corresponding comments were downloaded. Using PRAW, the following data was extracted for each comment: the exact date it was published, the author, the number of upvotes/downvotes and the content. This data was subsequently saved in an Excel file. At the end of the process, there were as many Excel files as there were posts analysed. Comments on posts published on a specific date could have been written on the same day or on subsequent days. Since the problem to be addressed was to check daily trends, it was necessary to collect all comments on the different posts in one place. To fulfil this task, a database was created using Access, allowing for easier queries, in particular by filtering the publication date of the comments.

The entire dataset consists of 791,588 comments, a number already filtered out by deleted, removed and BOT comments, and covers a period from 16-05-2021 to 11-06-2021. The total number of comments without filtering would have been 831,078.

CHAPTER 4 - Natural language processing and its applications

In accounting and finance, the availability of company reports, news articles and social media content provide a wealth of data on which to apply textual analysis. Textual analysis can be used to monitor this data even in real time to gain an informational advantage, before this information is read and assimilated by people. Stock market investor evaluations incorporate not only quantitative but also qualitative data. Textual analysis can also be called qualitative analysis and is less precise than the quantitative methods generally used in accounting and finance. Despite this imprecision, the use of textual analysis nevertheless represents an opportunity to be exploited. Textual analysis resides in many disciplines including natural language processing (henceforth NLP) and its applications including information extraction and sentiment analysis. Qualitative data to be used as input requires translation into quantitative data (Loughran and McDonald 2016).

Text analysis requires a large and solid text corpus as a prerequisite. Several text corpora form textual corpora, which collect many documents in natural language, written or spoken, linked together. Textual corpora are often stored in electronic form, while historical corpora are converted into electronic form so that they can be analysed easily. The size of a corpus can vary from small to large, ranging from tens to hundreds of gigabytes. Corpora may consist of unstructured data in a single language or in several languages. Text corpora are used for textual analysis. To use a text corpus, pre-processing must be carried out so that it is ready for subsequent analysis.

4.1 - Natural Language Processing and Natural Language Toolkit

The language that everyone commonly uses to communicate in both verbal and written form still represents a very important source of data. Despite its high potential this source is still not fully exploited. Human languages such as English, Italian, Spanish and all the many others are natural languages. A natural language has matured, expanded and mutated by humans through use and communication and has been created to be understood by other people. In contrast, a computer programming language is artificially created and constructed. So, there is a dichotomy between the two types of language. Textual data very often refer to natural language in the form of a word, sentence, or document. This is unstructured data, but it follows a precise syntax and semantics and is the basis for text analysis and NLP. In order for the computer to be able to analyse unstructured data, as it cannot directly employ mathematical or statistical models, NLP and algorithms must be used to transform textual data into numerical data and thus become readable by the computer. This conversion process is called 'vectorization' or 'word embedding'.

In contrast to unstructured data, structured or semi-structured data, being characterised by fields or markup make it easier for the computer to process them. The analysis of unstructured textual data represents an important possibility for researchers and organisations to derive valuable and useful information in the scientific, economic, social and cultural fields. In order for the computer to process and understand this data, NLP must be used to obtain useful outputs. NLP falls into the fields of computer science, engineering, Artificial Intelligence (AI) and computational linguistics. The interplay of these fields makes it possible to design applications and systems that link machine and natural language, i.e. Human-Computer Interaction (HCI). In short, we can say that NLP is about the computer manipulation of natural language. In recent years, NLP is playing an increasingly important role. As far as the corporate field is concerned, it is of importance that an ever-widening number of people have practical knowledge of NLP (Chen, Chiang and Storey 2012; Sagiroglu and Sinanc 2013).

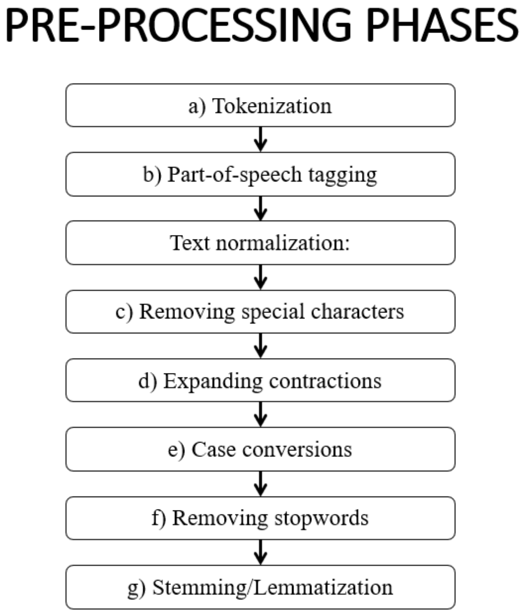
An important ready-to-use resource for NLP analysis is the Natural Language Tool-Kit (henceforth NLTK) as it contains corpora, lexical and grammatical resources, algorithms and models for natural language processing. NLTK was initially developed in 2001 at the University of Pennsylvania and then extended by academic experts over the years. NLTK frameworks allow Python programmers to process and analyse textual data, even in different languages, saving them a lot of time and effort that they would have had to spend on creating the code to perform these operations. Thus, NLTK allows researchers to focus their work on algorithms for solving real problems. Because of these characteristics, NLTK is functional for NLP activities and allows important research projects to be carried out (Bird, Klein and Loper 2009).

4.2 - Text pre-processing

In order to be able to analyse text data, which are initially in a raw format, they must be cleaned, normalised and pre-processed in such a way that the processor can recognise them. Pre-processing, or text processing, consists of a series of steps to accomplish this task. Thorough and solid pre-processing of the text is crucial for NLP and text analysis as the derived textual data, be it words, phrases, sentences, or tokens, constitute the initial input elements that will be augmented later to proceed to more complex analyses, such as information extraction and sentiment analysis. The famous IT motto 'garbage in, garbage out' is of utmost importance because if the text is not processed accurately and correctly, the results obtained from the analysis will not be satisfactory and relevant.

The following figure (Figure 4.1) represents the most common techniques of text pre-processing based on *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data* written by D. Sarkar in 2016.

Figure 4.1: Pre-processing phases.


PRE-PROCESSING PHASES
a) Tokenization
b) Part-of-speech tagging
Text normalization:
c) Removing special characters
d) Expanding contractions
e) Case conversions
f) Removing stopwords
g) Stemming/Lemmatization

Source: Sarkar 2016.

a) Tokenization

Knowing the structure of natural language in detail is useful in text analysis. A text document can be subdivided into several parts following a precise hierarchy: sentence → clause → phrase → word. Tokenisation is the process of breaking down and subdividing the textual corpus into smaller meaningful elements called tokens. Tokens are independent textual elements that can represent words, but also special characters and punctuation. Specifically, the token is a string of encoded bytes representing the text. Tokenisation of the textual corpus may be carried out before or after the elimination of potentially unnecessary characters and symbols in the textual data. The decision of when to carry out tokenisation depends on the type of problem and the data being addressed. Tokenisation is a fundamental process in the various pre-processing stages, especially in text normalisation where stemming and lemmatisation operate on the individual word.

NLTK provides several useful options for tokenizing: word_tokenize, TreebankWordTokenizer and RegexpTokenizer. NLTK recommends using the word tokenize function, which is the default option. The word tokenize function splits standard contractions, e.g. 'don't' is transformed into 'do' and 'n't', and punctuation handling, e.g. commas, full stops followed by whitespace, as separate tokens. The end result of tokenization separating the raw text into tokens is the starting point for tagging and normalising them, with the goal of having clean, standardised data for using NLP.

b) Part-of-speech tagging

Once one is able to access tokens extracted from textual documents, the next step after tokenization in the normal NLP process is tagging, also called part-of-speech tagging or POS tagging. In natural language, the same word can have different functions and therefore it is important to be able to distinguish between them. For example, the word 'work' can be either a verb or a noun. POS are lexical categories to which words belong according to their context and syntactic role. The main lexical categories to which words can belong are:

- N(oun): Nouns refer to things, people, places, or concepts, for example, table, child, Italy, kindness. Nouns can appear after adjectives and determiner and can be the subject or object of the verb. N represents simplified noun tags for common nouns like table and NP for proper nouns like Italy.
- V(erb): Verbs express events and actions, such as sing and walk.
- Adj(ective): Adjectives describe nouns and can be used as predicates (for instance, heavy in the sofa is heavy) or as modifiers (for example, heavy in the heavy sofa). English adjectives can have an internal structure (for example, increase+ing in the increasing prices).
- Adv(erb): Adverbs transform verbs to specify the manner, time, direction, or place of the event expressed by the verb (for example, slowly in the prices increased slowly). Adverbs can also modify adjectives (for example very in the road is very long).

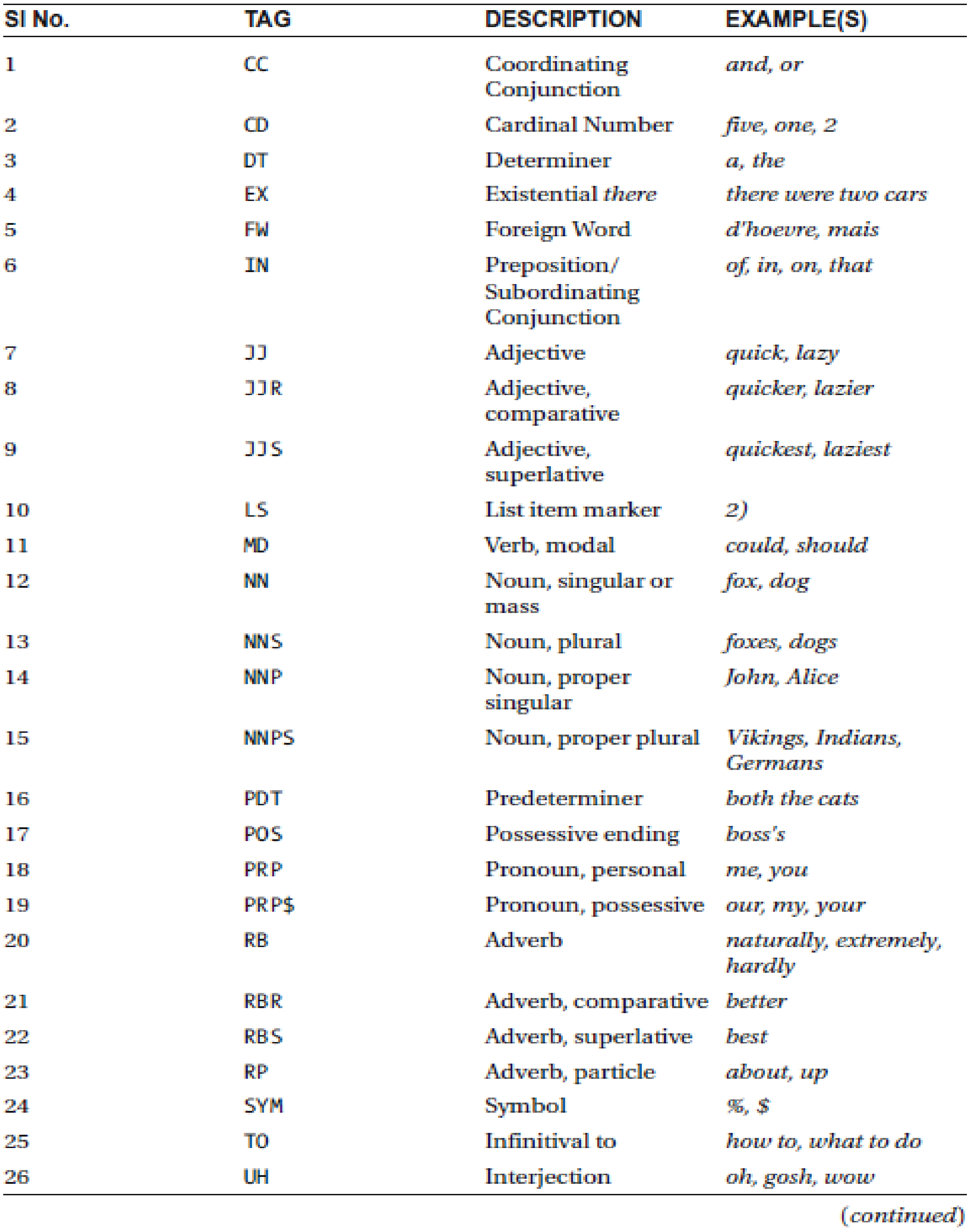
The N, V, ADJ and ADV tags are open lexical categories, i.e. composed of words included in an open vocabulary. One of the characteristics of natural language is that it evolves over time by adding new words. Open lexical categories, as the name suggests, allow these new words to be added to the vocabulary. Closed lexical categories consist of a finite number of words and do not admit new additions. These categories include pronouns, prepositions, articles/determinants, modals and personal pronouns.

Tagging classifies and labels each token with the appropriate POS tag. POS tags encode how the token functions in context. Tagging is particularly useful in NLP applications because it allows this valuable information to be exploited for specific analyses, such as identifying which names occur most frequently and disambiguating the meaning of words.

A POS tagger processes a series of words and assigns the relevant POS tag to each of them. The POS tagger recommended by NLTK is the pos tag which is based on the Penn Treebank tagset, one of the most widely used in various text analysis and NLP applications. The Penn Treebank is a project that originated in the late 1980s at the University of Pennsylvania. A tagset is a set of tags used for a specific problem. The Penn Treebank tagset consists of 36 parts of speech, structural tags, and tense markers (e.g. for verbs VB, for singular nouns NN, for adverbs RB, for adjectives JJ). In NLTK, the tagging process returns tuples (tags, tokens), where the tag is a case-sensitive string and the most likely in a given context. A tuple is an ordered sequence of elements similar to a list, but once created it is immutable.

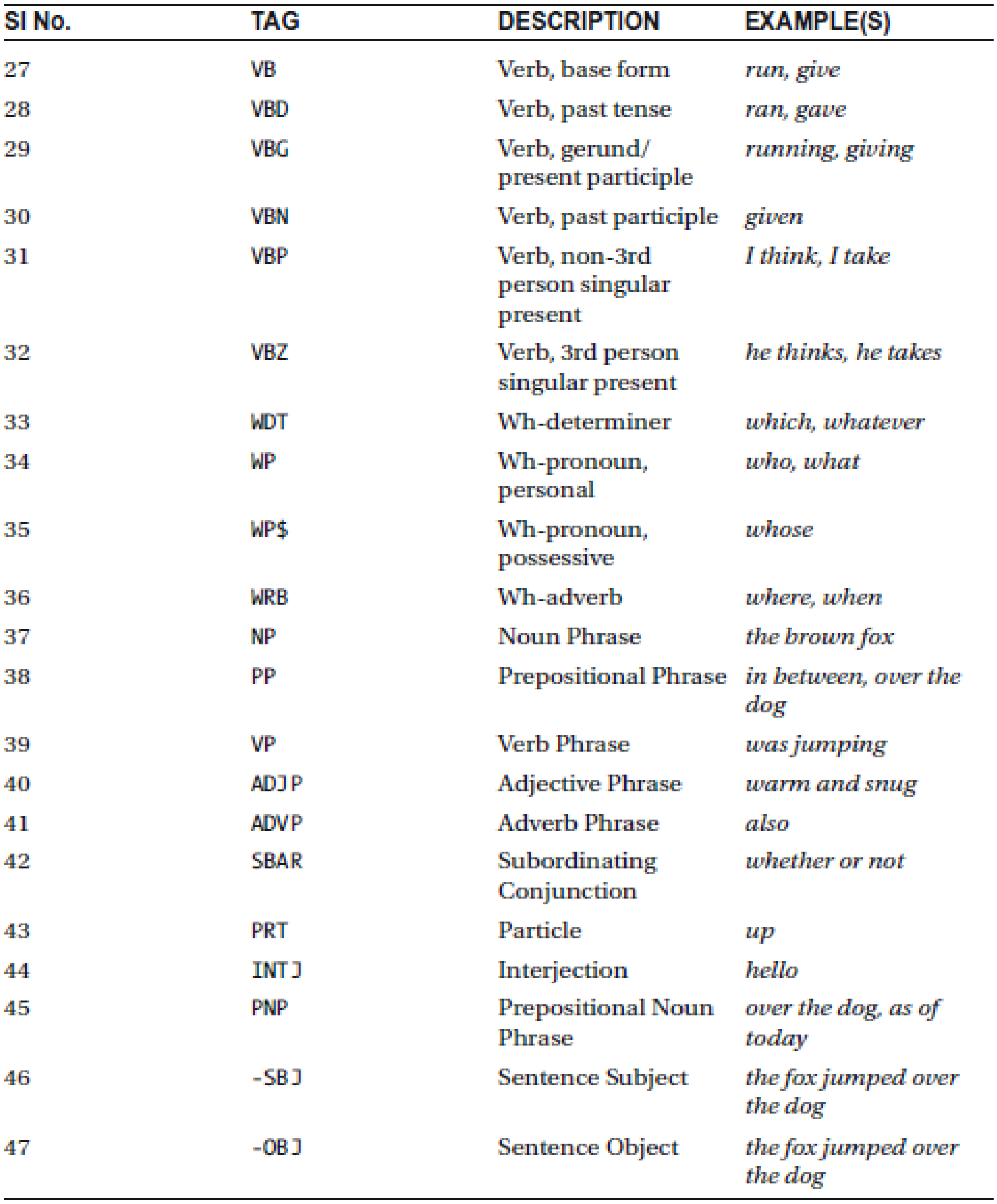
The figure below (Figure 4.2) details an overview of the Penn Treebank tagset and related examples.

Figure 4.2: Tagset Penn Treebank and related examples.



Sl No.	TAG	DESCRIPTION	EXAMPLE(S)
1	CC	Coordinating Conjunction	*and, or*
2	CD	Cardinal Number	*five, one, 2*
3	DT	Determiner	*a, the*
4	EX	Existential *there*	*there were two cars*
5	FW	Foreign Word	*d'hoevre, mais*
6	IN	Preposition/ Subordinating Conjunction	*of, in, on, that*
7	JJ	Adjective	*quick, lazy*
8	JJR	Adjective, comparative	*quicker, lazier*
9	JJS	Adjective, superlative	*quickest, laziest*
10	LS	List item marker	*2)*
11	MD	Verb, modal	*could, should*
12	NN	Noun, singular or mass	*fox, dog*
13	NNS	Noun, plural	*foxes, dogs*
14	NNP	Noun, proper singular	*John, Alice*
15	NNPS	Noun, proper plural	*Vikings, Indians, Germans*
16	PDT	Predeterminer	*both the cats*
17	POS	Possessive ending	*boss's*
18	PRP	Pronoun, personal	*me, you*
19	PRP$	Pronoun, possessive	*our, my, your*
20	RB	Adverb	*naturally, extremely, hardly*
21	RBR	Adverb, comparative	*better*
22	RBS	Adverb, superlative	*best*
23	RP	Adverb, particle	*about, up*
24	SYM	Symbol	*%, $*
25	TO	Infinitival to	*how to, what to do*
26	UH	Interjection	*oh, gosh, wow*

(*continued*)



SI No.	TAG	DESCRIPTION	EXAMPLE(S)
27	VB	Verb, base form	*run, give*
28	VBD	Verb, past tense	*ran, gave*
29	VBG	Verb, gerund/ present participle	*running, giving*
30	VBN	Verb, past participle	*given*
31	VBP	Verb, non-3rd person singular present	*I think, I take*
32	VBZ	Verb, 3rd person singular present	*he thinks, he takes*
33	WDT	Wh-determiner	*which, whatever*
34	WP	Wh-pronoun, personal	*who, what*
35	WP$	Wh-pronoun, possessive	*whose*
36	WRB	Wh-adverb	*where, when*
37	NP	Noun Phrase	*the brown fox*
38	PP	Prepositional Phrase	*in between, over the dog*
39	VP	Verb Phrase	*was jumping*
40	ADJP	Adjective Phrase	*warm and snug*
41	ADVP	Adverb Phrase	*also*
42	SBAR	Subordinating Conjunction	*whether or not*
43	PRT	Particle	*up*
44	INTJ	Interjection	*hello*
45	PNP	Prepositional Noun Phrase	*over the dog, as of today*
46	-SBJ	Sentence Subject	*the fox jumped over the dog*
47	-OBJ	Sentence Object	*the fox jumped over the dog*

Source: Sarkar 2016.

Text normalization

Text normalization, also called text cleansing or wrangling, is a pre-processing step and consists of organising, cleaning and standardising text data for use by analysis and NLP applications as input. The goal of text normalization is to reduce word variations to a common representation of them. Tokenization can also be part of text normalization.

c) Removing special characters

One of the relevant steps in text normalization is the removal of punctuation and special characters that are often used in sentences and are useless in text analysis. This technique can be carried out before or after tokenization. Depending on the problem being addressed, apostrophes can be retained in sentences and contracted words can be expanded.

d) Expanding contractions

Contractions are words or syllables in abbreviated versions created by removing specific letters and sounds. They can be in written or spoken form. In formal language contractions are rarely used, while in informal language they are used more. Contractions are formed with an apostrophe character expressing two or more words and this poses a problem in text analysis and NLP when tokenizing and standardising words. To solve this problem, contractions can be expanded in the text. Often in the English language, these contractions occur by removing one of the vowels in the word, such as *don't* becomes *do not* or *aren't* becomes *are not.*

e) Case conversions

To standardise sentences or words, it can be very useful to change their case. The two most commonly used methods are the uppercase conversion, where all letters become uppercase, and the lowercase conversion, where all letters become lowercase. In addition, there are other methods such as the sentence case, where only the first letter of the first word of a sentence becomes uppercase, the proper case, where the first letter of all words in a sentence becomes uppercase, and the title case or headline case, which is similar to the proper case with the difference that minor words remain lowercase.

f) Removing stop words

Stop words, such as *a, the, you*, for example, are often removed from the text during normalisation as insignificant, while only words with more meaning are retained. In the corpus of the text, stop words are the words that appear most frequently. Every language has its own stop words. There is no universal and complete set of them. NLTK has a set of stop words for the English language that can be used to eliminate all words in the text that correspond to the stop words in this set. Stop words can also include the negations *not* and *no*, but great care must be taken when eliminating them, as the actual context of the sentence may be lost, especially if sentiment analysis is being performed. Therefore, one must carefully consider and balance which stop words to eliminate.

g.1) Stemming

Words are the smallest, independent units with their own meaning present in a language. Within a word, several morphemes can be present. Morphemes represent the smallest unit in natural language that has its own meaning, be it a word or part of a word. Morphemes are composed of stems and affixes. Affixes are added to the word stem, changing its meaning or creating a new one. There are several categories of affixes, such as prefixes, suffixes, simul fixes. Word stems represent the basic form of words to which affixes can be added to create new words. This process is called inflection. The reverse process of inflection is stemming, i.e. transforming the word from its inflected form to its basic form. For example, to the word *work*, adding affixes to it results in new words such as *works, worked, and working*. The word *work* is the word stem. By stemming the three example words into their inflected forms, one obtains their word stem. In this way, words are standardised, facilitating various applications such as information retrieval. Stemming is used extensively by search engines to provide more exact results to the user regardless of the form of the requested words.

NLTK provides several stemmers:

- The Porter Stemmer, developed by Dr. Martin Porter, is one of the most used and has five stages for stemming, each of which has its own set of rules.
- The Lancaster Stemmer, also called Paice/Husk Stemmer, has over 120 rules for stemming.
- The SnowballStemmer stems in 13 different languages in addition to English.
- The RegexpStemmer allows you to build your own Stemmer by applying rules defined by the user.

The choice of which type of stemmer to use can vary depending on the problem you are dealing with.

g.2) Lemmatization

The lexeme includes all the different forms of a word, e.g. *singing, sang, sings*.

The lemma is the basic form of one or more words, net of their inflections or variations, as found in the dictionary. Going back to the example above, the lemma of the lexeme *singing, sang, sings* is *sing*.

Lemmatization is slightly different from stemming. Both remove affixes from the word to obtain its basic form. The difference between the two processes is that stemming can return the root stem which is not necessarily a lexicographically correct word and may not be found in the dictionary; whereas lemmatisation only returns the root word, or lemma, which can always be found in the dictionary. Lemmatisation is much more time-consuming than stemming as it removes affixes from words if and only if the lemma is found in the dictionary.

NLTK uses WordNet to perform lemmatization. The WordNet, created at Princeton University in 1985, is a lexical database for the English language consisting of words, their definitions, relations and examples, and the set of synonyms (synsets). The WordNetLemmatizer class uses the Morphy() function to find the lemma of a word, using the word and the part of speech, comparing them with the WordNet corpus; it then removes the affixes from the word until it finds it in the WordNet itself. In case of a negative outcome, the initial word will be reported unchanged. Successful lemmatisation depends on the correctness of the part of speech.

In the previous paragraphs, a survey was made of the various techniques most commonly used to process, normalise and standardise text. In addition to those listed, there are also others such as the removal of numbers and the conversion of emojis. Once these techniques have been implemented, it is possible to analyse text data in more detail in order to perform more complex operations on them.

4.3 - Information extraction

The multitude of unstructured textual data contains an infinity of valuable information that is hidden. Man, with his limited resources in terms of time and attention, would hardly be able to extract knowledge from them given the huge number of documents (Cukier 2010). Information extraction captures and summarises the main topics, concepts, and key phrases in documents. This task is facilitated by information extraction techniques that allow meaningful insights to be derived to facilitate rapid decision-making. An initial analysis of the textual data involves the extraction of key words and phrases to enable a general understanding of the topics and entities present in them.

One of the possible approaches is the extraction of n-grams. This approach exploits the concept of n-grams to identify adjacent successions of tokens in a given fixed length interval n: unigrams, n=1 (single tokens) - bigrams, n=2 (tuple of tokens e.g. 'I,' 'ate') - trigrams, n=3 (tuple of tokens e.g. 'I,' 'ate,' 'well') and following. If a pattern with several n-grams is constructed, it will be more difficult to have the same adjacent words repeating because too much noise would be created, thus losing the purpose of the analysis. The choice of interval size is a trade-off between the sensitivity and specificity of the model to find the right balance.

Another possible approach, already discussed above, that can be exploited is the use of POS tagging. For example, it is possible to extract from a text the most frequent individual nouns, verbs or adjectives, or the most frequent combination of them, such as noun and verb, adjective and noun, and so on.

Extracted text data can be represented through visual tools such as word clouds. The words within them vary in colour and size according to the frequency of the words in the examined text. The larger the words, the more important and used they are in the text, and vice versa. The purpose of using word clouds is to summarise and visualise textual information in an appealing manner and to make it possible to interpret the results in a simple manner. It is a flexible tool in that it allows you to decide how many words to display and can also be customised in other aspects, e.g. the word cloud can take the form of a mask, the background colour, font and word colour can be changed, and so on. Word clouds can be used to represent n-grams and POS-tagging results.

4.4 - Sentiment analysis

In the contemporary world, always more people use social networks to express their ideas, feelings, judgements, preferences, opinions, comments on facts or experiences related to people, consumer goods, travel, work, medicine, finance, politics and more (Daas et al. 2015). In this way, a multitude of data is put online that, if properly processed, can provide very important information to understand people's tendencies in various fields (Manyika et al. 2011): for example, it could be used by a company to plan market strategies to obtain better results or to predict a certain financial trend.

Social media have revolutionised traditional means of collecting trend data in various fields. Previously, the most widely used means were classic surveys and questionnaires, now information can be extracted directly from the content expressed by users to supplement it (Conrad et al. 2021; Smith and Gustafson 2017). Economics, politics, business and other social spheres are and will increasingly be influenced by social media. The latter help to detect social sentiment and understand and exploit the mechanisms of influence. Sentiment analysis, also called *opinion mining*, is the process of determining the sentiment expressed by people across this multitude of textual data in a given time frame and context (Pang and Lee 2008). Generally, the dictionary-based method is used to perform this analysis.

Istat began to see in this type of analysis, the possibility of deriving interesting statistical data in particular on the Italian economic situation. In line with Eurostat and other national statistical institutes (Eurostat 2024), Istat used new data sources and methods to derive valuable new information. In this way, gaps in topics not yet covered can be filled quickly (Istat 2023a).

4.4.1 - Social Mood on Economy Index

Istat's Social Mood on Economy Index (SMEI) is an experimental index, started on 10 February 2016, that measures Italians' daily sentiment on the economy in real time by analysing public tweets, i.e. tweets that are visible even to those who do not have a tweeter account, written in Italian. Domain experts have developed a filter that extracts and processes only those messages that contain keywords or sets of keywords, eliminating useless or statistically irrelevant messages from the outset. For this reason, the filter excludes retweeted tweets, while retaining quoted and commented tweets. The filter for extracting tweets consists of 60 words or sets of keywords that were mostly chosen from the consumer confidence survey questionnaire. Compared to the latter, the SMEI measures a much broader phenomenon.

Towards the end of 2021, given the increasing development of sentiment analysis techniques, the SMEI was updated and implemented with an additional second-level filter containing 115 words or sets of keywords. The second-level filter does not add new words but refines them, i.e. declines them in more ways. By doing so, while decreasing the number of tweets extracted and analysed, there will be a greater skimming of irrelevant ones than those that actually talk about the economy.

Approximately 26,000 tweets are processed daily and after being cleaned and normalised are analysed using the sentiment analysis method. This method is based on an Italian-language sentiment lexicon, i.e. a vocabulary, to which pre-calculated positive and negative sentiment scores are associated. A clustering algorithm separates the tweets into three distinct classes, negative, neutral and positive. Then, using a given measure of central tendency to the distribution of the numerical sentiment values of the tweets, the daily index value is obtained. However, as there was no score for neutral sentiments in the vocabulary, they were not considered for the index calculation. With the updated vocabulary, which also scores neutral sentiments, it is no longer necessary to do the clustering operation, whereas the way of calculating the index does not change (Istat 2023b).

4.4.2 - Dictionary-based method

The dictionary-based method is one of the most commonly used text analysis methods to date (Gentzkow et al. 2019). This method is often employed in sentiment analysis and uses a very valuable tool called *word lists, lexicons or dictionaries*. The first step in applying this method is to choose which word list to use according to one's goal. A word list is a collection of words representing a certain attribute with common sentiment, such as positivity and negativity, so that it can be identified in the text. By counting the words related to their attribute in the word lists and dividing them by the total number of words in the text, a comparative measure of tone or sentiment can be derived. Thus, if more positive words are present in a text than negative words, an optimistic tone will result. In addition, some word lists represent other attributes besides positivity and negativity, such as uncertainty, pleasure and pain. The use of public dictionaries to measure sentiment makes it possible to avoid the subjectivity of the analyst, can be applied to large quantities of text documents, and can replicate and make it easier for other researchers to conduct analysis (Loughran and McDonald, 2016.).

Originally, the Harvard General Inquirer (GI) word lists were used in the fields of sociology and psychology. Despite its use in these fields, Harvard GI word lists have also been used in other fields such as accounting and finance to measure sentiment. Initially, most researchers in these fields, such as Tetlock (2007), Engelberg (2008), Tetlock (2008) used the Harvard GI word lists, as they were among the few that were readily available.

The sentiment measured through the use of Harvard IV-4 lists, specifically created for contexts other than accounting and finance, is criticized by Loughran and McDonald. Loughran and McDonald (2011) show that 73.8% of the negative words on the Harvard IV-4 word list generally do not correspond to the pessimism present in the financial context. For example, the words *tax, liability, capital and cost* in most 10-K documents are neutral in nature as they only describe business operations. Furthermore, they show that negative Harvard IV-4 words can misclassify the tone of words in specific industries, such as cancer for pharmaceutical industry and mine for precious metals and coal industries.

The Loughran-McDonald Master Dictionary was created based on the 2of12inf dictionary. The latter is part of the 12dicts. 12dicts is a collection of English word lists by Alan Beale derived from 12 dictionaries, including 8 ESL (English as a Second Language) dictionaries, 4 "desk dictionaries", and other sources. To create the 12dicts, the words included in the sources just described were collected and correlated. The main goal of 12dicts was to develop a basic English vocabulary. The 12dics lists are divided into 4 directories that contain lists with similar properties:

- American: lists with American English words.
- International: lists with American English and British English words.
- Lemmatized: lists that harmonize other lists organized in such a way as to clarify the relationships between words.
- Special: unique lists that do not belong to other directories.

The 12dicts lists are distinguished from other word lists that can be found on the web for several reasons: they include only common words, they are checked to minimize errors, most of them are public, one can count on a large number of lists with different characteristics in terms of size and type of words. Typically, only a single 12dicts list is used. The 2of12inf list contains over 80,000 words and is made up of inflections of words and no abbreviations, acronyms or names. All words contain at least two or more characters (Scowl 2016).

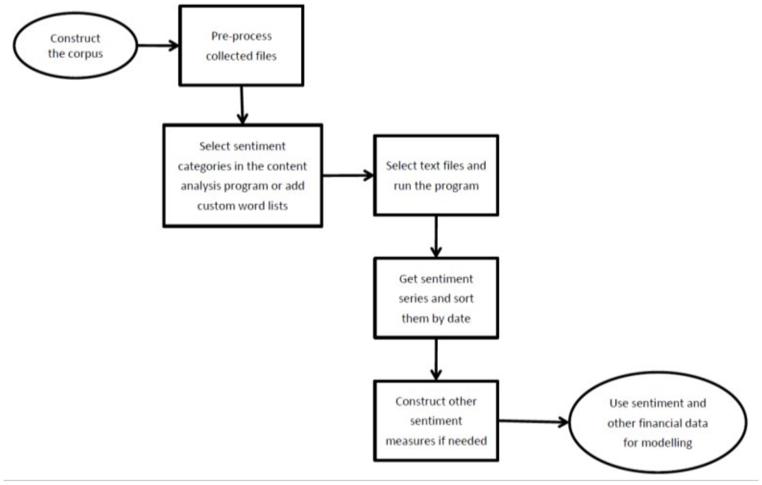
The Master Dictionary, created by Timothy C. Loughran and Bill McDonald, is based on the 2of12inf word list which is expanded using Electronic Data Gathering, Analysis, and Retrieval EDGAR 10-X filings (10-K, 10-K/A, 10-K405, 10-K405/A, 10KSB, 10KSB/A, 10-KSB, 10-KSB, 10-KSB/A, 10-KSB/A, 10-KSB, 10-KSB/A, 10-KSB/A, 10-KSB40, 10KSB40/A) from 1993 to the current year to be explicitly used to better reflect the tone of the language present in the financial documents. The Master Dictionary consists of six different lists of sentiment words: *negative, positive, uncertainty, litigious, strong modal, weak modal,* and *constraining*. Since natural language evolves, the Master Dictionary is updated every year by including new words through two steps: first, the words that have a consistent frequency are automatically extracted from the documents and subsequently, experts classify them by evaluating their most probable use (Loughran and McDonald 2022). Kearney and Liu (2014) state that the Loughran and McDonald word lists have been widely used in several more recent studies such as those by Jegadeesh and Wu (2012), Chen et al. (2013), Liu and McConnell (2013).

The use of the word lists developed by Loughran and McDonald (2011) in this research has limitations, making it unsuitable. Their word lists were created by analysing 10-Ks documents that use formal language. Therefore, they cannot accurately capture the sentiment contained in dynamic comments on social media, which are also characterized by slang, profanity, emojis, symbols and acronyms (Loughran and McDonald 2016). Loughran and McDonald (2020) suggest using lists of words more suited to the context of interest.

An alternative word list (Renault 2022) for extracting sentiment from comments in social media was created by Assistant Professor Thomas Renault (2017), resulting to be suitable for this research because it was developed on the StockTwits, very similar to Reddit. StockTwits provides valuable information source providing instant global data regarding the financial markets. StockTwits allows users to classify their message as “bullish” (positive sentiment) or “bearish” (negative sentiment). Renault exploited this feature to extract the most frequent relevant words to classify them as positive or negative, and to build a new public word list using machine learning techniques.

To sum up, the whole theoretical part (Figure 4.3) schematizes the different phases of the dictionary-based approach to extract sentiment from unstructured data. First of all, it is essential to create the text corpus consisting of a collection of unstructured data obtained from one or more sources such as company documents, media articles and online comments. Text corpus pre-processing is a necessary step to be able to analyse text data. Subsequently, an already existing word list is selected or a new personalized one can be created that best meets the purpose of the research. The sentiment values are obtained by applying the dictionary-based approach to the text and the results are grouped by date. Finally, sentiment measures along with other financial variables can be used for hypothesis testing and financial modelling (Cruz, Kinyua and Mutigwe 2023).

Figure 4.3: Sentiment extraction (dictionary-based approach).


Construct the corpus
Pre-process collected files
Select sentiment categories in the content analysis program or add custom word lists
Select text files and run the program
Get sentiment series and sort them by date
Construct other sentiment measures if needed
Use sentiment and other financial data for modelling

Source: Cruz, Kinyua and Mutigwe 2023.

SECTION II - EMPIRICAL RESEARCH AND RESULTS

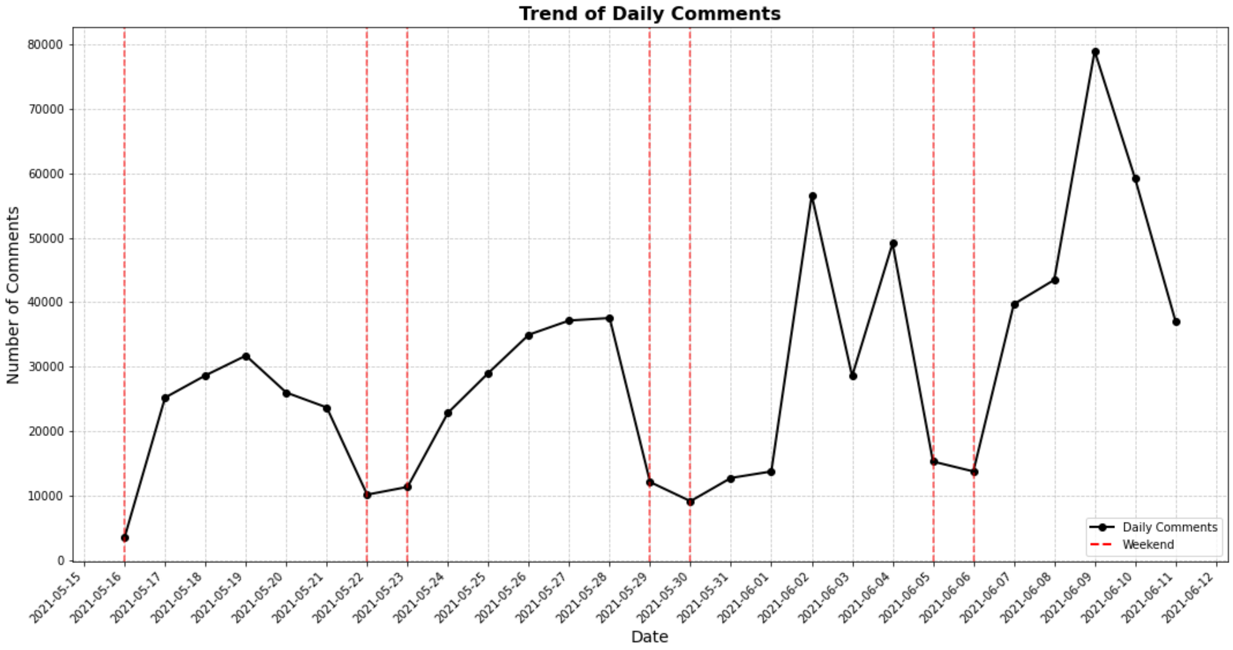
CHAPTER 5 - Analysis of the total dataset

5.1 - Number and length of comments and usage of emojis

Before carrying out more specific analyses, it is useful to know some statistics, such as the quantity and length of comments and the usage of emojis, regarding the total dataset on which one is working to identify any potential trends.

The line graph below (Figure 5.1 based on Table 1 data) shows the number of comments posted by users that make up the dataset spread daily. It is remarkable noticing that there are days when there are higher peaks and lower ones. The least minimum number of comments occurs during the weekend while the most outstanding and the most significant number of comments is concentrated between Monday and Friday, i.e., when the stock market is open. The highest peak is 78.927 comments and the average number of comments posted per day is 29.318. The period analysed is approximately one month and in this period of time one can notice a certain weekly seasonality in the number of comments posted by users. Some higher peaks are immediately noticeable and it is helpful to investigate and deepen them further to understand better the possible factors that determine them.

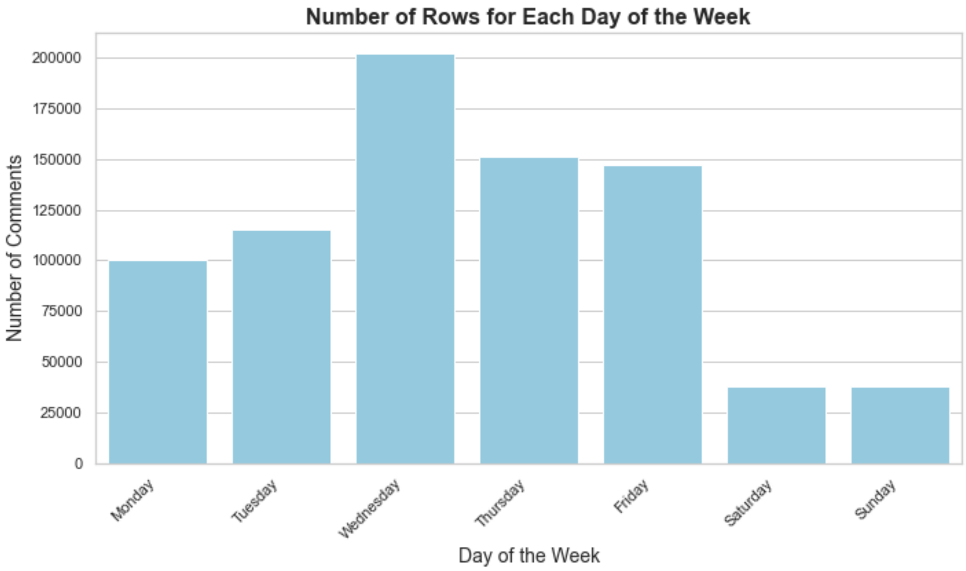
Figure 5.1: Number of comments in the dataset spread from 16/05/2021 to 11/06/2021.



Another interesting statistic is the number of comments grouped into the days that make up the week (Figure 5.2 based on Table 2 data). As previously mentioned, the weekend has the lowest number of comments, more precisely on Saturday and on Sunday, with 37.635 and 37.919, respectively, as it corresponds to the stock market closing days. The days with the highest number of comments are Wednesday with 202.198, Thursday with 150.905 and Friday with 147.433.

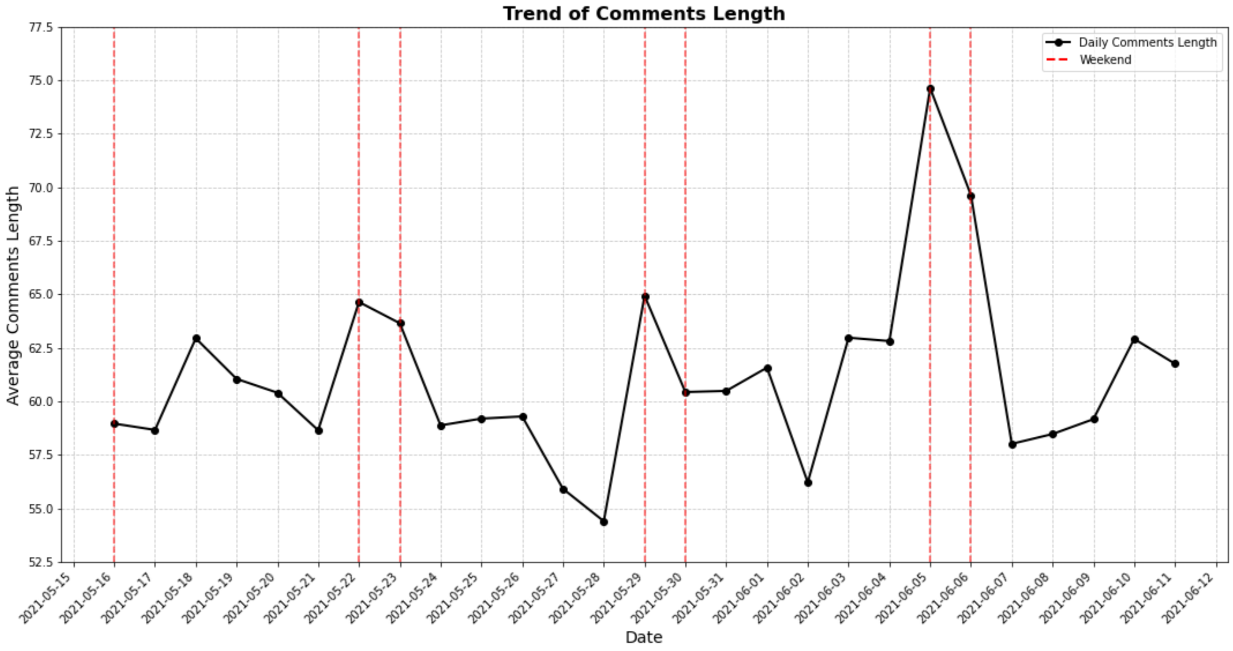
A statistic based on the closings of the Wall Street Stock Exchange over the last 95 years has been conducted, which is also confirmed in the shorter period of 10 years, highlighting that not all the days of the week have the same value in the Stock Market. It turns out that Monday is the least lively day during the closings while Wednesday is the most dynamic, demonstrating a strong seasonality on a weekly scale. One factor to be considered is that after the end of the weekend when the Stock Market opens, markets may face unexpected events that can have significant negative impacts (black and grey swan). This data is essential for both investors and traders, who can develop the best strategies to get more benefits (Lops 2023).

Figure 5.2: Number of comments grouped by days of the week from 16/05/2021 to 11/06/2021.



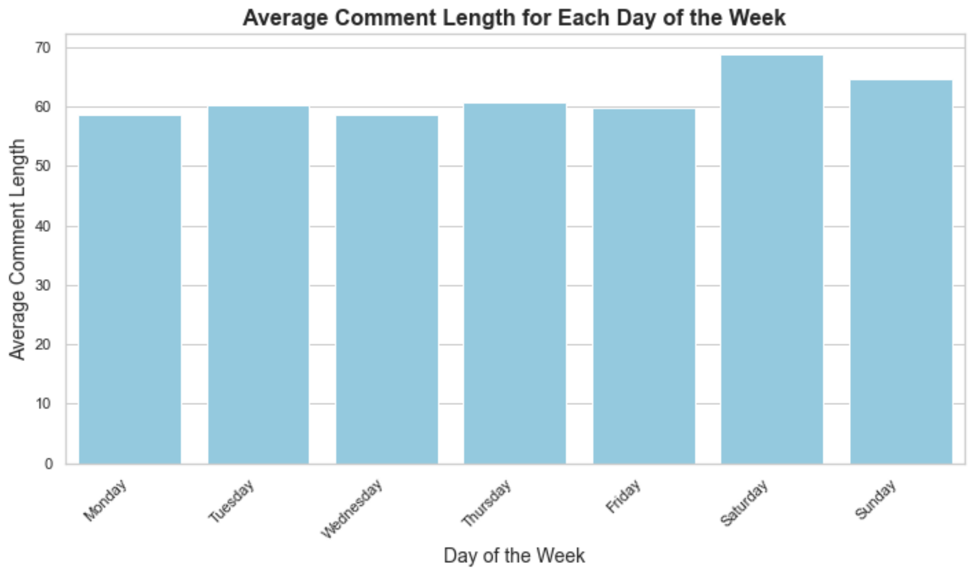
Another interesting statistic is to analyse whether the number of characters, namely the length of the comment, varies in any way compared to the results of the previous graph (Figure 5.2), which concerned the number of comments. It can be seen in the graph below (Figure 5.3 based on Table 1 data) that the most extended and most full-bodied comments occur during the weekend, while in the previous graph (Figure 5.2) it turns out that the smallest number of comments appears in the same period. It can be assumed that this is because users have more free time and are calmer and more relaxed during the weekend as the Exchange is closed. Shorter comments occur when the Stock Market is open and users are more frantic. Hence, one can hypothesize that there is an inverse relationship between the total number of comments and their length.

Figure 5.3: Length of comments in the dataset spread from 16/05/2021 to 11/06/2021.



The previous graph (Figure 5.3) shows that the most extended comments occur on the weekend. As seen from the graph below (Figure 5.4 based on Table 2 data), there is a sure consistency in the length of comments on the days when the Stock Market is open. The average length of comments from Monday to Friday is 59,6 characters, while on Saturday and Sunday, it is 68,8 and 64,6 characters, respectively.

Figure 5.4: Length of comments grouped by days of the week from 16/05/2021 to 11/06/2021.



Emojis are increasingly used in social media to facilitate digital communication globally. Emojis were born in Japan at the end of the 20th century and in the Japanese language, the word 'emoji' means 'picture character'. Emojis are Unicode graphic symbols that are used to express concepts and ideas. The number of users who are available to express emotions, feelings and moods by reinforcing and amplifying text messages is very high. Moreover, emojis capture the attention of message readers more quickly. The sentiment and meaning of emojis depend on the context in which they are used and evolve. Emoji sentiment analysis shows that emojis tend to occur more at the end of comments. Emojis can be exploited to extend text-only sentiment analysis (Novak et al., 2017). Considering the 791.588 comments in the analysed dataset, it appears that 86.466 comments contain at least one emoji, i.e. 10,92% of the total comments. Table 5.1 shows some examples of how these results were obtained. Comments with an emoji are assigned the value 1, while those in which it does not appear are assigned the value 0 (Attached 2). Even if there are multiple emojis, the variable's value always remains, so the simple sum represents the number of comments with emojis.

Table 5.1

I buy AAPL and MSFT!	0
I buy AAPL 🚀 and MSFT!	1
I sell MSFT and AAPL 🐻 🐻 💀.	1

In particular, from the emojis typically used in WSB (Appendix A), it appears that the emojis that reflect a positive sentiment are more significant than the negative ones, 45,317 and 11,395, respectively (Table 3). Their sum equals 56,712, corresponding to the total number of comments containing at least one emojis. Furthermore, this sum corresponds to 65.59% of the 86,466 comments containing all the possible emojis that can be used. This last data confirms that the sample of emojis used in this analysis is very accurate and reliable, although it takes into consideration only a part of the immense vastness of emojis available. Table 5.2 shows some examples of how these results were obtained. First, the comment emojis are converted into their name. The names of the most used emojis in WSB (Appendix A) are contained in a positive word list (for example, rocket) and a negative one (for example, skull and bear). The emojis are identified when a match occurs between the name of the emojis in the list and the one present in the comments. If there are several different emojis belonging to the sample in the same comment, they are all detected, but if the same emoji is repeated several times in a comment, it is only considered once.

Table 5.2

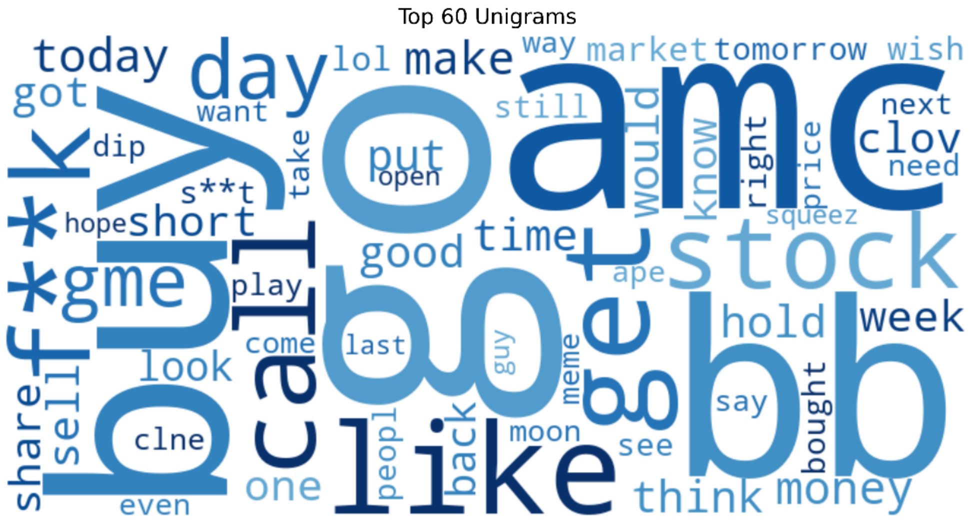
I buy AAPL and MSFT!	[]
I buy AAPL 🚀 and MSFT!	['rocket']
I sell MSFT and AAPL 🐻 🐻 💀.	['skull', 'bear']

5.2 - N-grams and POS-Tagging

Analysing the words that appear most frequently in the total dataset can provide exciting clues and summarize and highlight the themes most present in the users' comments. To carry out this type of analysis, it is important that the pre-processing, which cleans, normalizes and standardizes the text, has already been performed. This phase is necessary to bring the words back from their different inflections to their basic form and therefore, be able to be grouped together.

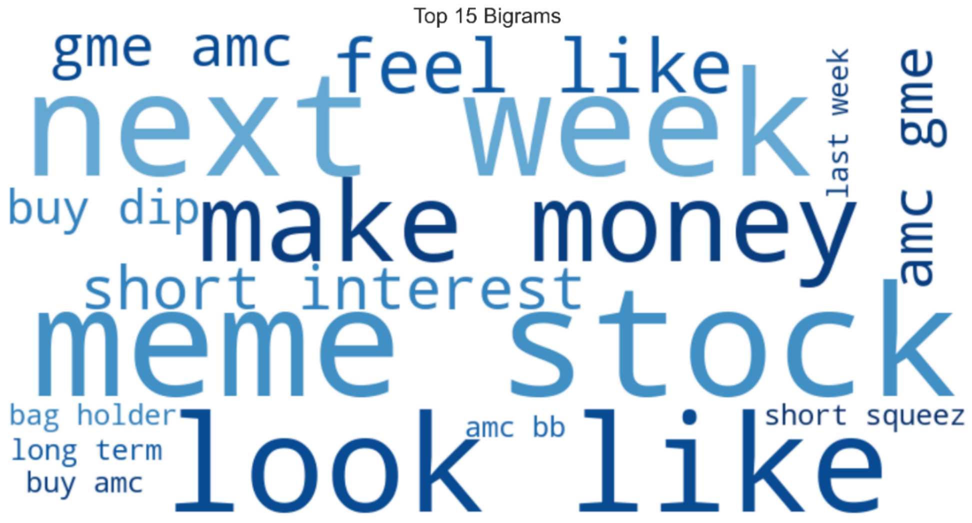
As can be seen from the word cloud (Figure 5.5), which includes the sixty most frequent unigrams, some tickers of listed companies such as AMC, BB, GME, CLOV, CLNE clearly emerge. The first five unigrams and their frequency are: "go" 74.864, "amc" 57.825, "bb" 48.276, "buy" 48.055 and "like" 47.316.

Figure 5.5: The 60 most frequent unigrams from 16/05/2021 to 11/06/2021.



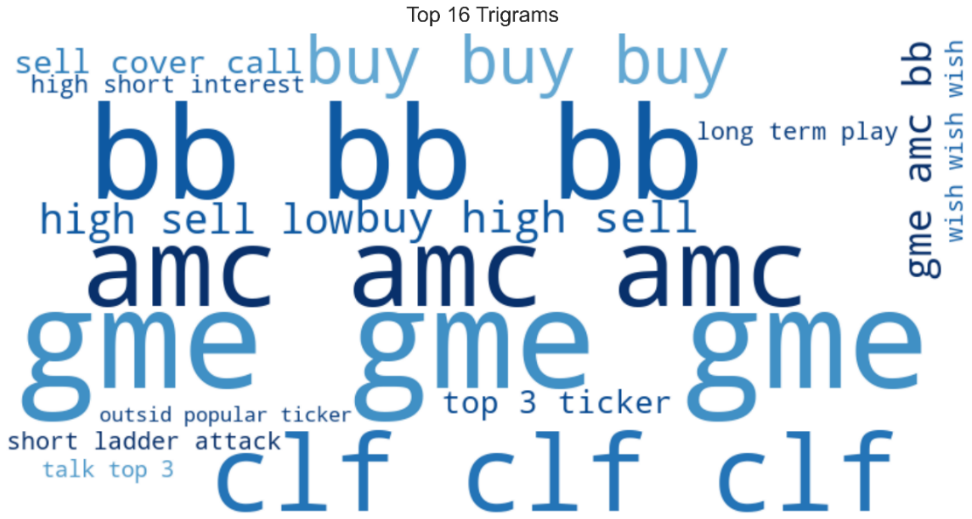
Analysing only a single word or unigram may not fully and comprehensively demonstrate the concept expressed in the users' comments. In the following word cloud (Figure 5.6), the fifteen most frequent bigrams are considered. The presence in pairs of some companies indicated in the previous word cloud (Figure 5.5) is confirmed. The first ten bigrams and their frequency are: "meme stock" 5.985, "next week" 5.336, "look like" 4.522, "make money" 4.323, "feel like" 3.096, "short interest" 2.858, "gme amc" 2.654, "amc gme" 2.601, "buy dip" 2.541 and "short squeeze" 2.515.

Figure 5.6: The 15 most frequent bigrams from 16/05/2021 to 11/06/2021.



In the following word cloud (Figure 5.7), the sixteen most frequent trigrams are considered. It is interesting to note the trigrams "talk top 3" and "top 3 tickers" which could explain the repeated presence of the corporate tickers AMC, BB, GME, their combination, CLF and WISH. The first five trigrams and their frequency are: "gme gme gme" 787, "bb bb bb" 686, "amc amc amc" 484, "clf clf clf" 439 and "buy buy buy" 426.

Figure 5.7: The 16 most frequent trigrams from 16/05/2021 to 11/06/2021.

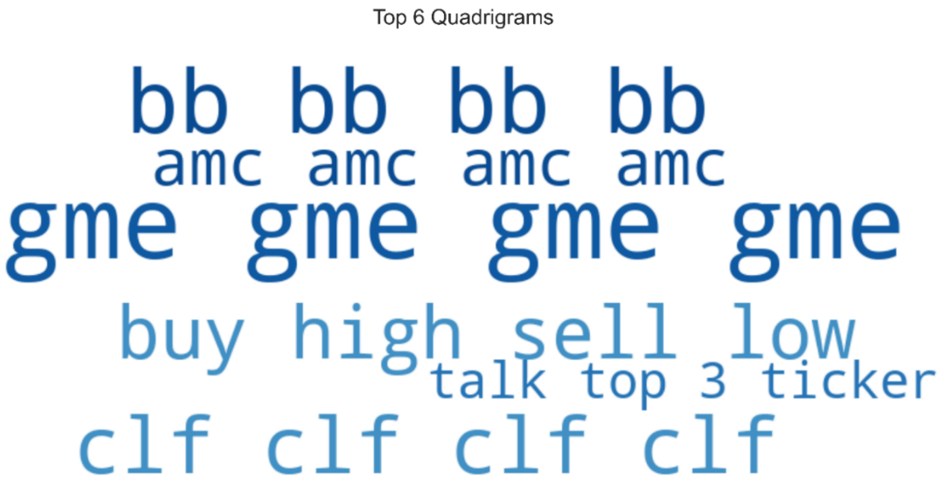


In the following word cloud (Figure 5.8), the six most frequent quadrigrams are considered. The presence of some companies indicated in the previous word clouds (Figure 5.5, 5.6, 5.7) is confirmed. The repetition of the ticker of the same company suggests some involvement and conviction of the users who place in them. The first six quadrigrams and their frequency are: “gme gme gme gme” 682, “bb bb bb bb” 513, “clf clf clf clf” 406, “buy high sell low” 349, “amc amc amc amc” 279 and “talk top 3 tickers” 260.

As the interval of the n-grams increases, the frequency of occurrences decreases because it is more difficult for the exact words to be repeated in the same order in different users’ comments.

In conclusion, the recurring presence of these companies expresses the interest of a large number of users over the period taken into consideration. It is also worth remembering that this evidence is present in general discussions in the financial sector, but despite this, users' interests still converge mainly on a small group of listed companies.

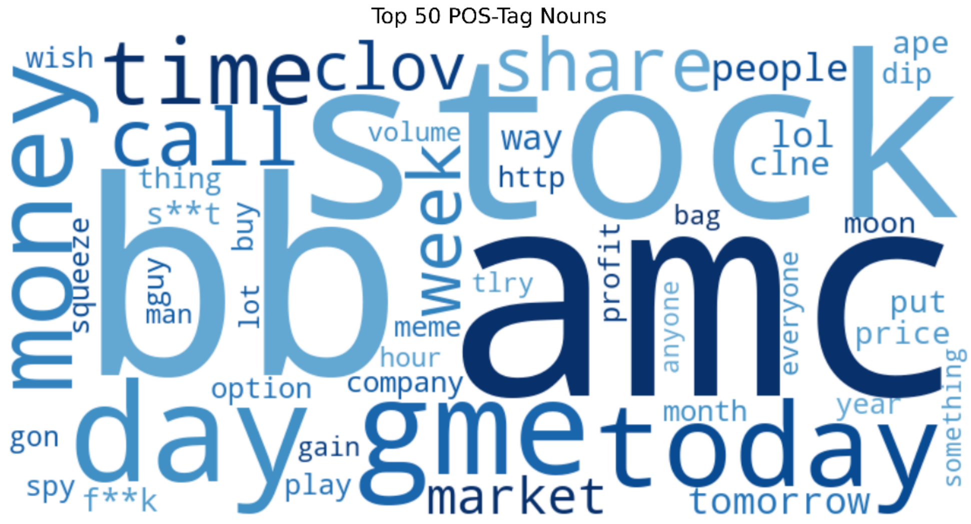
Figure 5.8: The 6 most frequent quadrigrams from 16/05/2021 to 11/06/2021.



The most frequent n-grams show the words that occur most often in the text without considering which of the various function words can take on within a sentence.

To identify what function a particular specific word performs it is essential to take into consideration the part of speech through post-tagging. In the following word cloud (Figure 5.9), the fifty most frequent POS-Tag nouns are considered. The top five nouns and their frequency are: “amc” 53.243, “bb” 46.393, “stock” 33.867, “day” 28.690, “gme” 27.406.

Figure 5.9: The 50 most frequent POS-Tag nouns from 16/05/2021 to 11/06/2021.



If one compares the first five results of the unigrams in Figure 5.5 (go, amc, bb, buy and like) with those of the nouns just highlighted (Figure 5.9), one notices that there are some differences, namely that no lexical categories, such as verbs, are no longer present while the tickers of the most mentioned listed companies still appear.

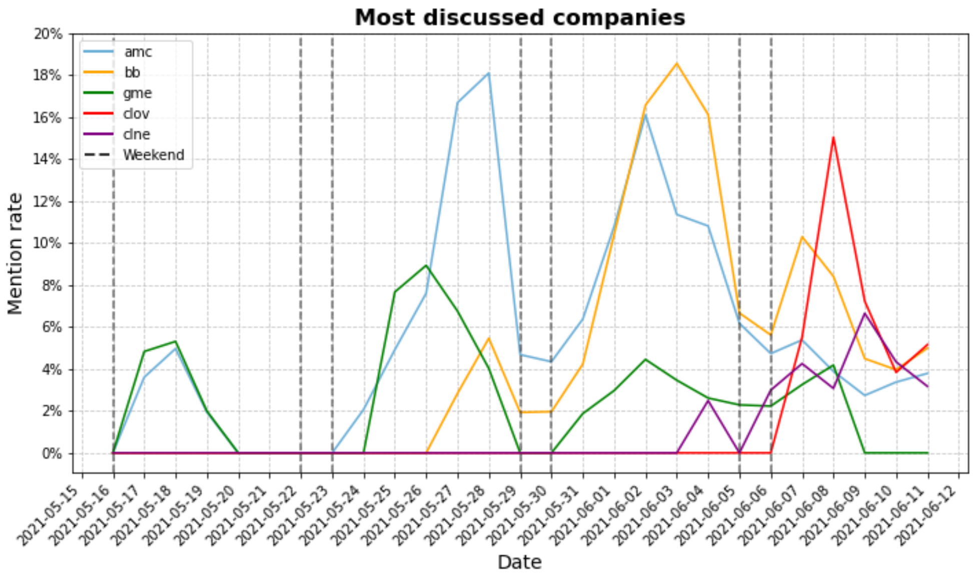
CHAPTER 6 - Daily analysis of the dataset

6.1 - POS-Tagging breakdown

Up to now, analyses have been carried out on the total dataset entailing exciting results but all at an aggregate level. From now on, the investigation focuses on a daily level to better understand how the large quantity of tickers of listed companies can be distributed over the reference period.

An initial search can be carried out by trying to break down the results of the POS-Tagging applied to the total dataset. The occurrences of each noun were counted and only the fifteen most mentioned nouns from the first day were considered. This process was repeated daily until the last day of the period under consideration was reached. The choice to consider only the fifteen most mentioned nouns was made to verify and confirm the consistency of the results over time. From the time series of the fifteen names, only the first five most mentioned tickers of the entire period examined and their respective number of daily occurrences were considered. Since each day has a different number of comments, it has been divided the number of occurrences of each ticker on the days it was mentioned by the number of comments on the given day. As can be seen from the graph below (Figure 6.1 based on Table 4 data), there were several peaks in the tickers mentioned during the period under consideration. The highest peaks occurred on days when the Stock Market was open while the lowest peaks corresponded to the days it was closed.

Figure 6.1: The tickers among the 15 most frequent nouns from 16/05/2021 to 11/06/2021.



6.2 - Group identity

After having noticed consistent peaks in the number of mentions of the prominent leading companies in the comments, it is interesting to investigate whether this phenomenon may also be associated with a growing aggregation and cohesion among the redditors.

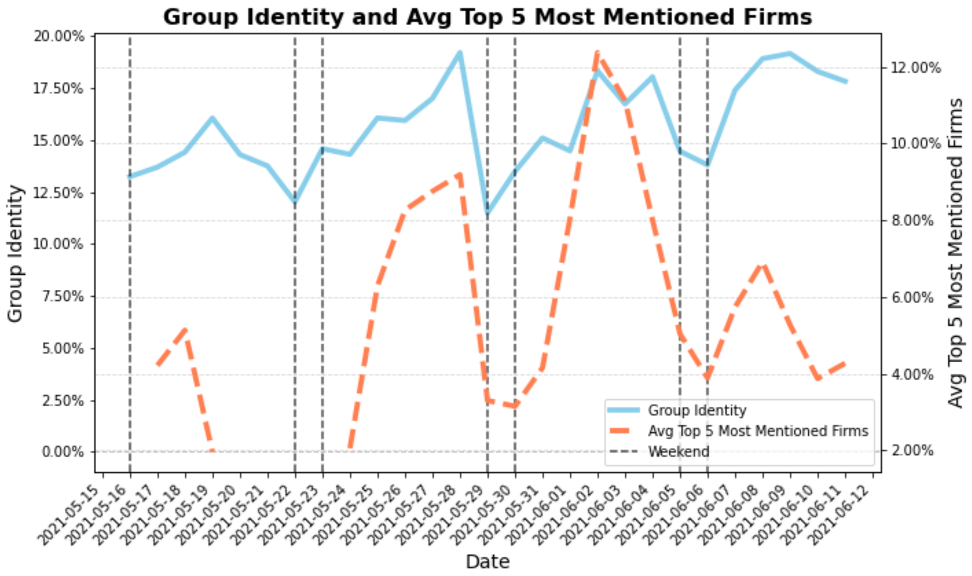
To do this, analysing words in comments can provide insights into users' psychological states. The words we use every day reflect what people pay attention to. For example, personal pronouns express the subject of such attention. A prevalence of the use of the first-person singular indicates a situation of individuality, while a prevalent use of the first-person plural indicates a situation of belonging and therefore can be an indicator that expresses group identity (Tausczik and Pennebaker 2010).

In this research, group identity was measured through the following index calculated on the comments of each day of the period considered (Lucchini et al. 2021):

$$\text{Group identity} = \frac{\text{we} + \text{our} + \text{ours}}{\text{we} + \text{our}+ \text{ours} + \text{I} + \text{me} + \text{mine}} \times 100$$

From the graph below (Figure 6.2 based on Table 4 data), one can see that there is a strong group identity which has a similar trend to that of the average without considering the zeros of the results of the five tickers among the fifteen most mentioned nouns in the previous graph (Figure 6.1). The analysis of the specific words that make up the group identity index reveals the increase in the sense of belonging to the group of redditors as the days go by and their possible consequent organization to carry out collective actions.

Figure 6.2: Comparison of the group identity index and the average of the top 5 most mentioned firms between the 15 nouns from 16/05/2021 to 11/06/2021.



6.3 - Most discussed listed companies and their stock prices

The investigation process continues by looking at the number of times some of the many companies that media and newspapers believed had a significant role in the performance of stock prices during the period under review were mentioned in comments. To do this, a filter of keywords was used, such as the ticker and company name. It has been used a dummy variable that took the value of 1 if keywords were found; otherwise, it took on the value 0. At this point, the values in the column of the dummy variable which were in the DataFrame were added together to obtain the number of times the keywords occurred. This process was repeated for each day of the period considered. Since not every day had the same number of comments, it was necessary to calculate the percentage of occurrence by dividing the number of times of mentions per day by the total number of comments per day. Comparing the percentages of daily occurrences with the share prices of the most mentioned companies in the dataset, it can be noted that the increase in mentions corresponds to an increase in the share price. On the contrary, when mentions start to decrease, the price also decreases (Figure 6.3 to 6.7 based on Table 5 to 9 data). Given the encouraging interim results based on the five most mentioned tickers, the analysis can be extended to a wider range of listed companies mentioned in the comments.

Figure 6.3: Comparison of AMC mention rate and stock price from 16/05/2021 to 11/06/2021.

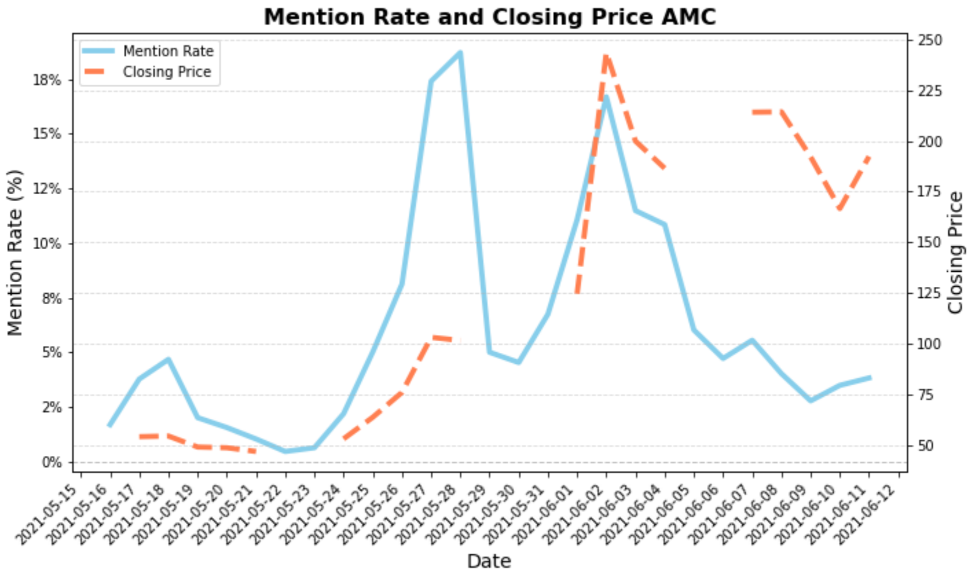


Figure 6.4: Comparison of BB mention rate and stock price from 16/05/2021 to 11/06/2021.

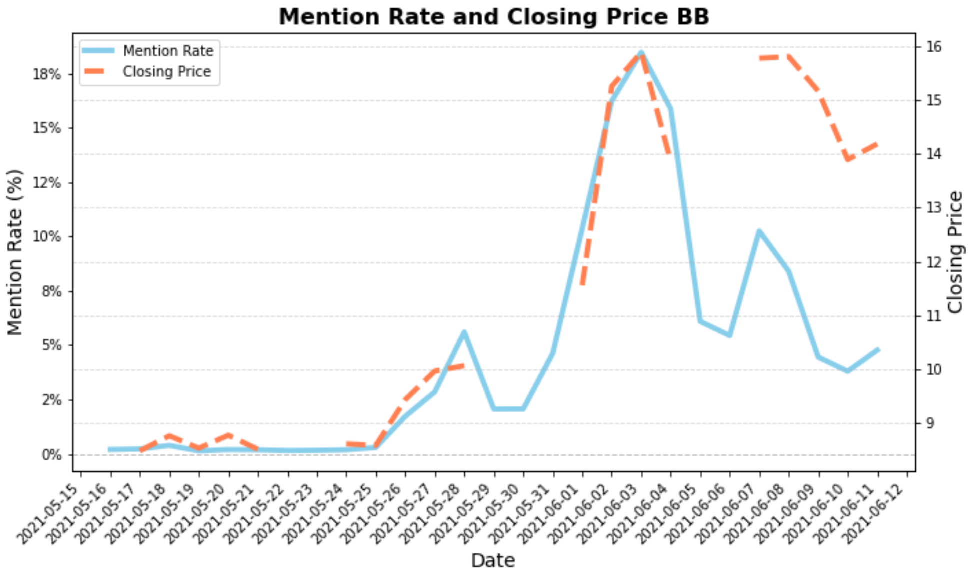


Figure 6.5: Comparison of GME mention rate and stock price from 16/05/2021 to 11/06/2021.

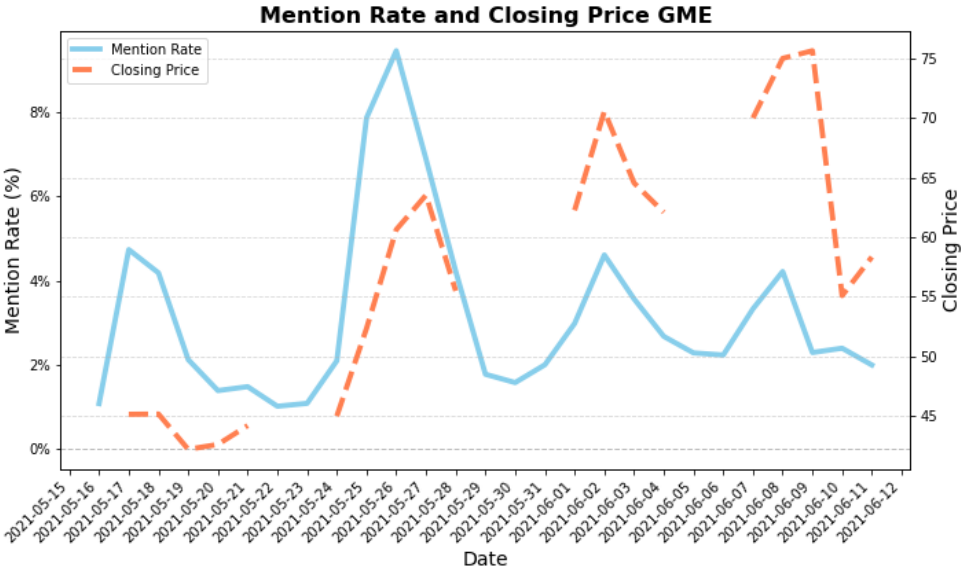


Figure 6.6: Comparison of CLOV mention rate and stock price from 16/05/2021 to 11/06/2021.

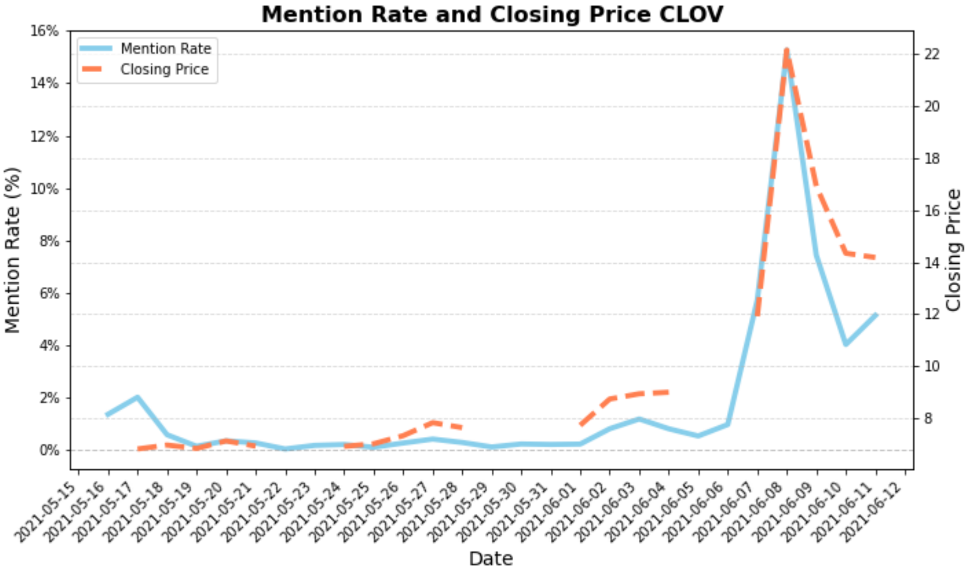
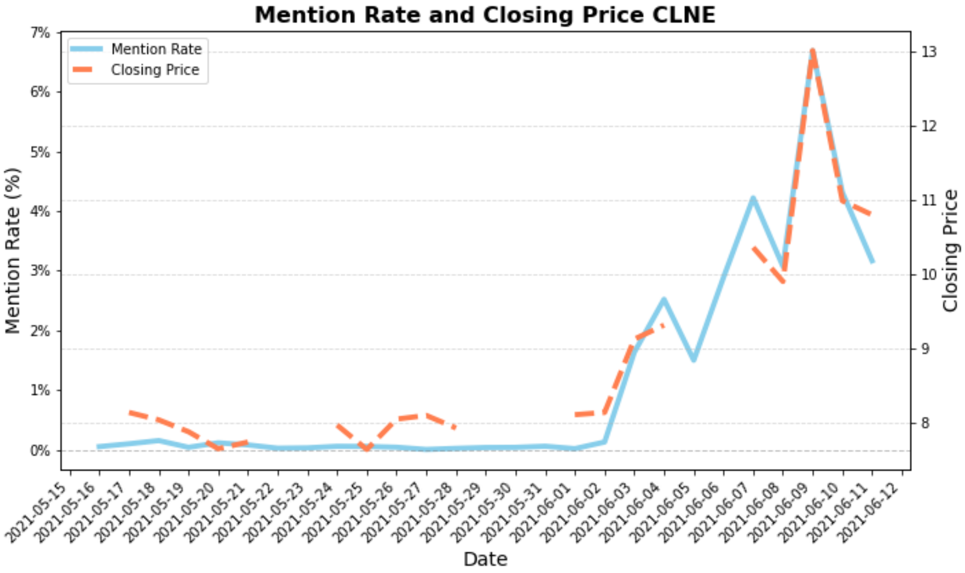


Figure 6.7: Comparison of CLNE mention rate and stock price from 16/05/2021 to 11/06/2021.



6.4 - Extension of ticker analysis

So far, the analysis has focused only on the most cited companies. To figure out which other companies were actually mentioned a different search has been carried out. The Refinitiv Eikon database gives the possibility to obtain various financial and non-financial data. One can download tickers related to listed companies on a specific stock exchange. In this research, the attention was focused on the NYSE and the NASDAQ since the companies most frequently mentioned in the analysed comments are listed on these exchanges and also because most of the Reddit users are American. By obtaining the list of companies and their tickers listed on the NYSE and NASDAQ, it was possible to search and count how many times the various tickers in the comments had been mentioned on a given day. By repeating the process each day, it was possible to obtain data that covered the entire period considered. By adding up all the occurrences of each company's ticker, it was possible to calculate the number of times the ticker was mentioned. The final result is a list (Table 11) that shows in descending order which tickers were mentioned the most and which the least during the period considered. This thesis takes into consideration only those companies extracted from comments which reached a total of mentions equal to or greater than 500 since they correspond to the minimum threshold of some lesser-known companies reported by online business newspapers as they had had sudden price changes such as MicroVision, Wendy's, Bionano Genomics, FuboTV (La Monica 2021; Miao and Stevens 2021; White 2021; Duprey 2021). Therefore, just because of the general nature of discussions it was possible to carry out an analysis that covers a vast range of stocks, accepting the challenge of future research by Cruz, Kinyua and Mutigwe (2023). This further information made it possible to test the hypothesis covered by this thesis using the linear regression model most commonly used by researchers for this type of task (Kearney and Liu 2014). Normally, the dependent variable subject to this type of analysis is a firm-level or market-level performance measure, such as future earnings changes (Li 2006), future returns on assets (Davis et al. 2012). The hypothesis to be tested is whether mentions, sentiment (calculated in Chapter 7 through Renault's word list) and group identity influence the stock returns of the most mentioned tickers. The return of each stock was calculated daily using the following formula:

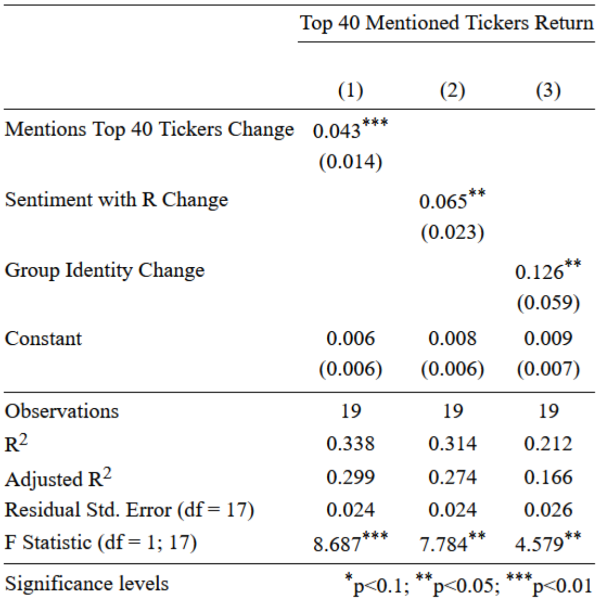
$$Rate\ of\ return = \frac{(Price\ today - Price\ yesterday)}{Price\ yesterday} \times 100$$

Once this is done, the return of the stocks with more than 500 mentions was calculated daily by taking the average of their daily returns. Subsequently, an investigation is carried out on the possible linear relationship between the dependent variable of the returns of stocks with more than 500 mentions and the independent variables of the variations in mentions, sentiment and group identity from 16th May 2021 to 11th June 2021 (Figure 6.8 based on Table 10 data). First of all, F Statistic is statistically significant with regard both to the first model because the p-value is lower than 0.01 and the second and third because it is lower than 0.05. One can therefore proceed with analysis of the individual models. Since the three models have only one independent variable, I took in consideration the R^2 (indication of the strength of the linear relationship between the dependent and independent variable) which is moderate but substantially good because there is a limited number of observations due to the change in the Reddit's API policy concerning the extraction of comments (Isaac 2023). Furthermore, in the models I am not considering all the independent variables that would be necessary to adequately explain the model. From the three models it can be stated that:

- Model (1): an increase of one percentage point in the mentions of company tickers produces an average increase of 0.043 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.
- Model (2): an increase of one percentage point in redditors' sentiment produces an average increase of 0.065 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.

- Model (3): an increase of one percentage point in the redditors' group identity produces an average increase of 0.126 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.

Figure 6.8: Linear regressions between stocks return and mentions, sentiment, group identity change.



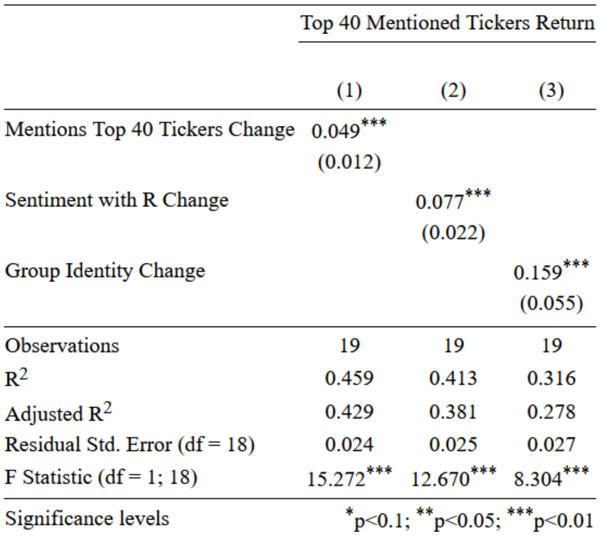
	Top 40 Mentioned Tickers Return		
	(1)	(2)	(3)
Mentions Top 40 Tickers Change	0.043***		
	(0.014)		
Sentiment with R Change		0.065**	
		(0.023)	
Group Identity Change			0.126**
			(0.059)
Constant	0.006	0.008	0.009
	(0.006)	(0.006)	(0.007)
Observations	19	19	19
R^2	0.338	0.314	0.212
Adjusted R^2	0.299	0.274	0.166
Residual Std. Error (df = 17)	0.024	0.024	0.026
F Statistic (df = 1; 17)	8.687***	7.784**	4.579**
Significance levels	*p<0.1; **p<0.05; ***p<0.01		

Given that the constant in all models is not statistically significant, you can proceed to estimate the regression models without constants (Figure 6.9 based on Table 10 data). First of all, F Statistic turns out statistically significant for the first, second and third models because the p-value is lower than 0.01. Even so, you can proceed analysing the three models. It is important to highlight that R^2 can also be evaluated in models without intercept/constant. However, the interpretation of this parameter cannot be compared between models with and without intercept/constant. Therefore, the models without constant/intercept only, they appear to have a moderate R^2 in this case too. Then, in the process of assessing the individual parameters, it can be inferred that these are all statistically significant. From the three models it can be stated that:

- Model (1): an increase of one percentage point in mentions of company tickers produces an average increase of 0.049 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.

- Model (2): an increase of one percentage point in redditors' sentiment produces an average increase of 0.077 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.
- Model (3): an increase of one percentage point in the redditors' group identity produces an average increase of 0.159 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.

Figure 6.9: Linear regressions without constants between stocks return and mentions, sentiment, group identity change.



	Top 40 Mentioned Tickers Return		
	(1)	(2)	(3)
Mentions Top 40 Tickers Change	0.049^{***}		
	(0.012)		
Sentiment with R Change		0.077^{***}	
		(0.022)	
Group Identity Change			0.159^{***}
			(0.055)
Observations	19	19	19
R^2	0.459	0.413	0.316
Adjusted R^2	0.429	0.381	0.278
Residual Std. Error (df = 18)	0.024	0.025	0.027
F Statistic (df = 1; 18)	15.272^{***}	12.670^{***}	8.304^{***}
Significance levels	*p<0.1; **p<0.05; ***p<0.01		

Therefore, it can be claimed that on average all the independent variables analysed have a positive effect on the dependent variable. Such assessments, which emerged by evaluating the linear regression models, confirm what has been analysed throughout the entire research.

CHAPTER 7 - Sentiment analysis

The investigation continued by carrying out sentiment analysis with which it was possible to extract the sentiment from users' comments. Among the potential methods for this type of analysis, the dictionary-based approach was used in this research, which is also the most employed for this purpose. A dictionary classifies every word contained within it into a category which can be positive or negative. Furthermore, some dictionaries can have other categories such as uncertainty. The dictionary-based approach calculates sentiment by counting the number of positive and negative words present in a comment, allowing the comment itself to be labelled as positive, negative or neutral based on the prevailing result. The choice of which dictionary to use in the analysis is crucial to obtain reliable results. Table 7.1 shows some examples of how sentiment analysis is calculated with the classic dictionary-based approach. Assuming a word list of positive words that includes the word 'buy' and a word list of negative words that consists of the word 'sell', one can quickly notice that the presence of a negative particle preceding the words 'buy' and 'sell' ' leads to a misclassification of the sentiment as it is not really considered.

Table 7.1

I buy amc stocks 🚀 .	1	Positive
I didn't buy amc stocks 🚀 .	1	Positive
I sell amc stocks 🚀 .	-1	Negative
I didn't sell amc stocks 🚀 .	-1	Negative

However, in this research the classic dictionary-based approach was extended by also managing the most common possible negative 'not' and 'no' particles that precede the individual words present in the comments labelled with post-tagging as verbs by one or two positions. The management of these negative particles through the code used (Attached 3) makes a verb negative that would otherwise have been classified as positive and vice versa (Polanyi and Zaenen 2006, Asmi and Ishaya 2012). Table 7.2 shows the same examples reported in Table 7.1 calculating and improving the sentiment analysis using an adapted version of the classic dictionary-based approach that manages negations. One can quickly notice that the sentiment calculation now varies based on the negative particles. It is underlined that although the word 'not' is not present in its entirety, the expansion of the contractions in the pre-processing phase transforms 'didn't' into 'did' and 'not', thus managing to detect and consider it.

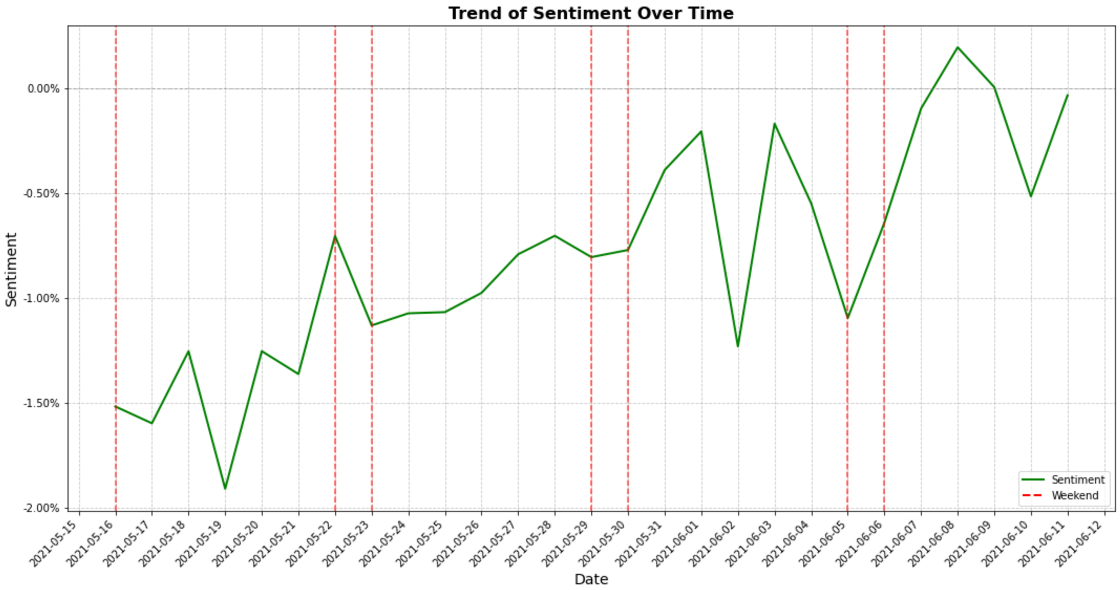
Table 7.2

I buy amc stocks 🚀.	1	Positive
I didn’t buy amc stocks 🚀.	-1	Negative
I sell amc stocks 🚀.	-1	Negative
I didn’t sell amc stocks 🚀.	1	Positive

7.1 - Use of the Master Dictionary Loughran – McDonald

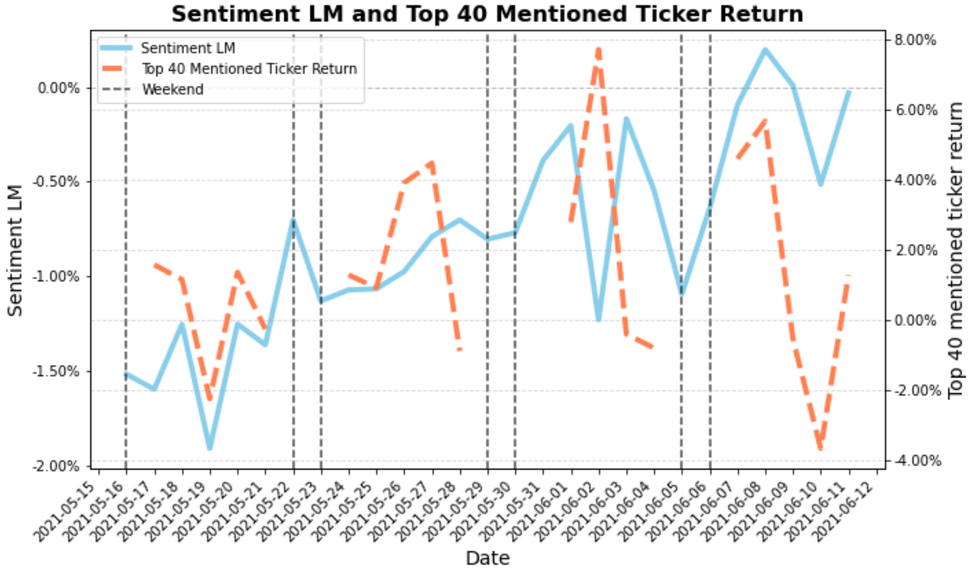
Initially, the Master Dictionary developed by Timothy C. Loughran and Bill McDonald, hereafter LM, was used in this research because it was created specifically for the financial context. The graph below (Figure 7.1 based on Table 12 data) shows the result. As can be seen, the average between positive and negative sentiment is always negative. This is due to the fact that in the LM there are more negative words than positive ones.

Figure 7.1: Sentiment using the Master Dictionary LM spread from 16/05/2021 to 11/06/2021.



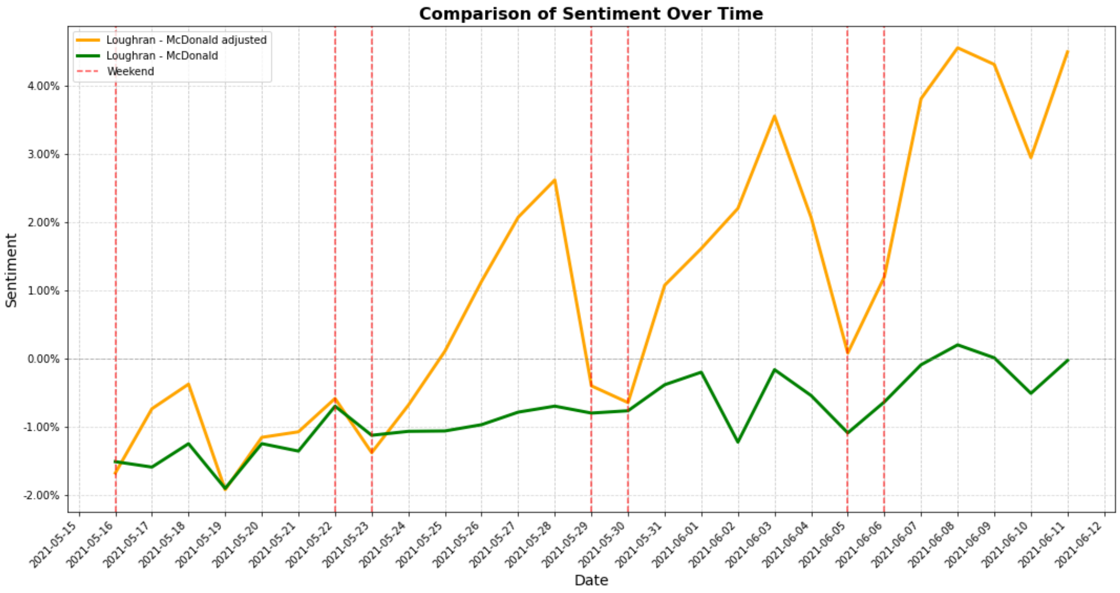
Among the many companies that were discussed in the comments on the dataset, the investigation focused on trying to relate sentiment to the stock returns of the most mentioned companies. In the graph below (Figure 7.2 based on Table 10 and 12 data), one can see that the sentiment resulting from LM has a similar trend to the stock returns of the forty most mentioned tickers, but fails to capture the highest peak of the returns of 2nd June 2021.

Figure 7.2: Comparison of the sentiment with LM and the return of the 40 most mentioned tickers from 16/05/2021 to 11/06/2021.



The use of LM in this research has limitations making it not entirely suitable, as it is based on a formal language. As a result, it is unable to accurately capture the sentiment contained in comments on social media, which are also characterized by slang, profanity, acronyms, symbols and emojis. The same authors suggest using a dictionary more suited to the context of interest. Trying to add the emojis and jargon (Appendix A and B) most used in WSB to the LM, in the graph below (Figure 7.3 based on Table 12 data) it turns out that the average between positive and negative sentiment is no longer just negative as it was in the previous graph (Figure 7.1) but presents a more positive average, demonstrating how emojis and jargon influence the result.

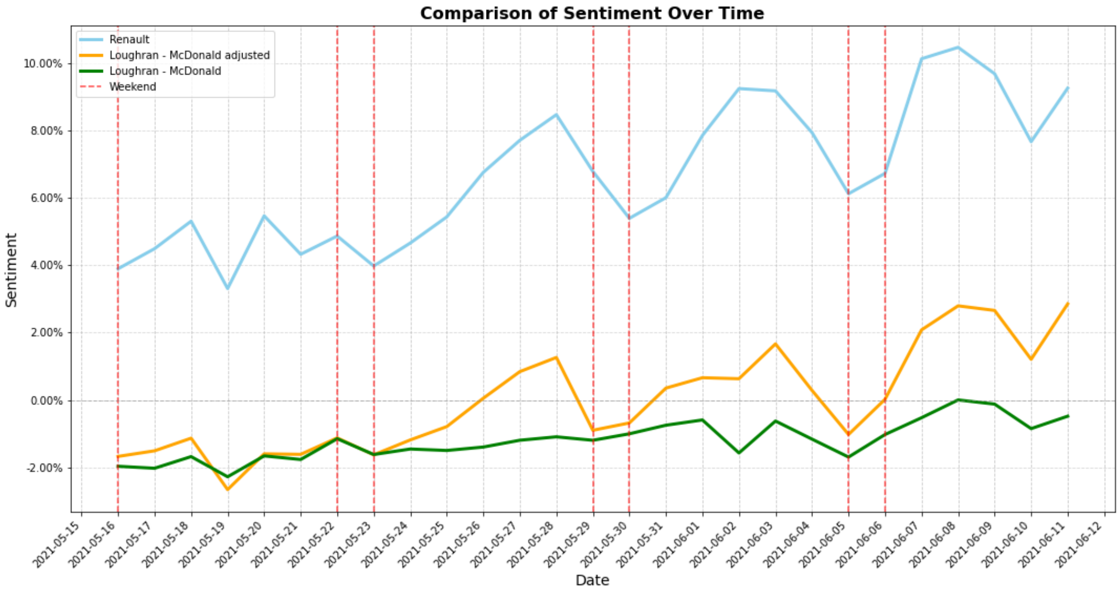
Figure 7.3: Comparison of the sentiment with adjusted LM and LM from 16/05/2021 to 11/06/2021.



7.2 - Use of the StockTwits Lexicon Renault

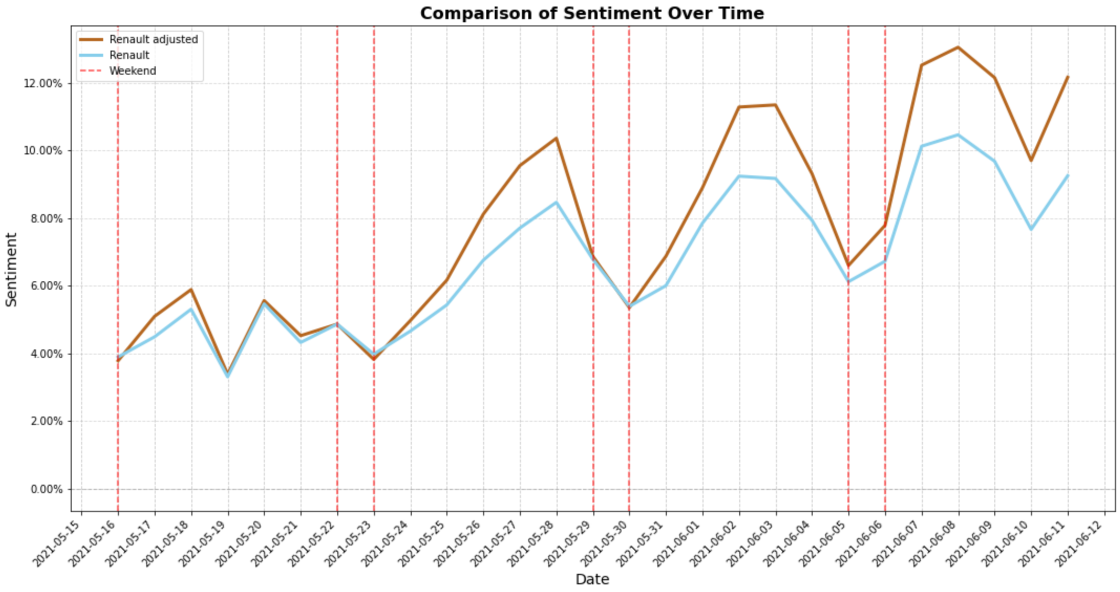
An alternative dictionary for extracting sentiment from comments in social media was created by Assistant Professor Thomas Renault, hereafter R. This is more adequate than those previously used in this research, because it was developed on the StockTwits social network, which is very similar to Reddit. R contains the same number of positive and negative words, solving the problem of imbalance present in the LM. As can be seen from the graph below (Figure 7.4 based on Table 12 data), with R the average of the sentiment between positive and negative is almost always positive, unlike the initial LM in which it is always negative (Figure 7.1). Instead, the average of the adjusted LM has a similar trend to that of R, although it is not always positive and has lower values.

Figure 7.4: Comparison of the sentiment with R, adjusted LM and LM from 16/05/2021 to 11/06/2021.



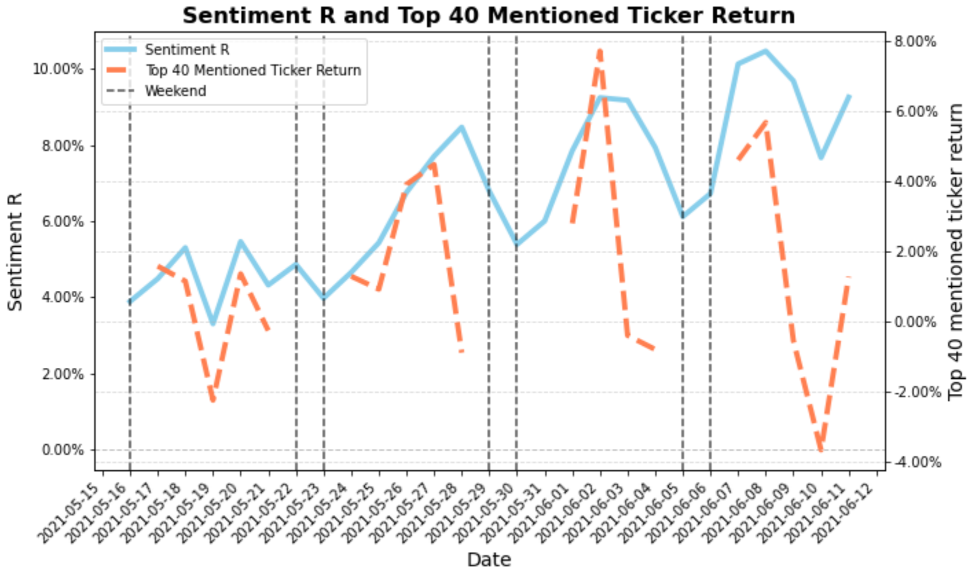
Trying to add the emojis and jargon (Appendix A and B) most used in WSB to R, one can see in the graph below (Figure 7.5 based on Table 12 data) that the trend of sentiment is the same in both versions albeit with higher values with adjusted R in correspondence with the mostly positive peaks. This indicates that the use of R turns out to be very valid for the purpose of this research as it is able to intercept users' sentiment in a more accurate manner even without the modification I made. For this reason and for a question of reliability, only R will be used.

Figure 7.5: Comparison of the sentiment with adjusted R and R from 16/05/2021 to 11/06/2021.



In the graph below (Figure 7.6 based on Table 10 and 12 data), one can see that the sentiment resulting from R has a similar trend to the stock returns of the 40 most mentioned tickers. Unlike the previous graph (Figure 7.2), in this case it can be seen that the sentiment is able to intercept the highest peak of the returns of 2nd June 2021.

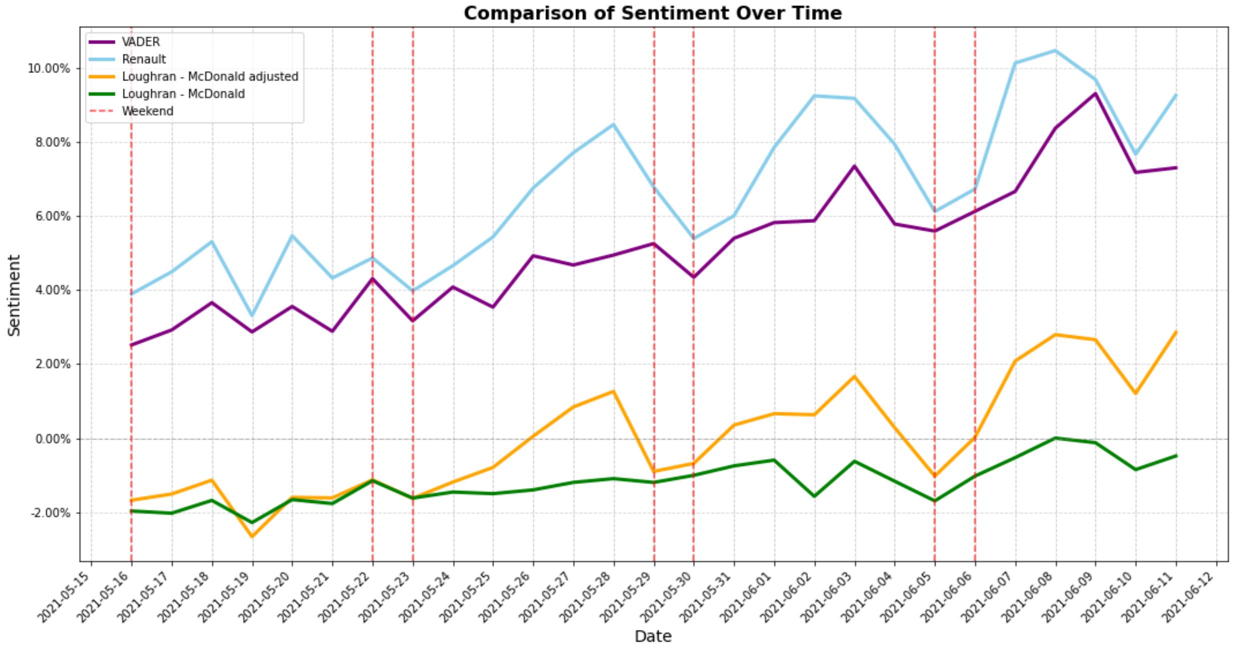
Figure 7.6: Comparison of the sentiment with R and the return of the 40 most mentioned tickers from 16/05/2021 to 11/06/2021.



7.3 - Use of VADER

Another essential critical and quick tool for performing sentiment analysis is VADER, acronym for Valence Aware Dictionary and Sentiment Reasoner, developed by C.J. Hutto and E. Gilbert. VADER uses a set of grammatical rules and a pre-built sentiment lexicon that includes words and phrases to which scores have been assigned that allow the positivity or negativity of the sentiment to be assessed. VADER is particularly suited to the social media context, where users express their thoughts in a short and informal way. This tool is also able to consider emojis (Hutto and Gilbert 2014). As can be seen from the graph below (Figure 7.10 based on Table 12 data), with the VADER the sentiment is always positive as it is using R and partly also adjusted LM, thus validating the trends of the previous graphs.

Figure 7.9: Comparison of the sentiment with VADER and the other dictionaries from 16/05/2021 to 11/06/2021.



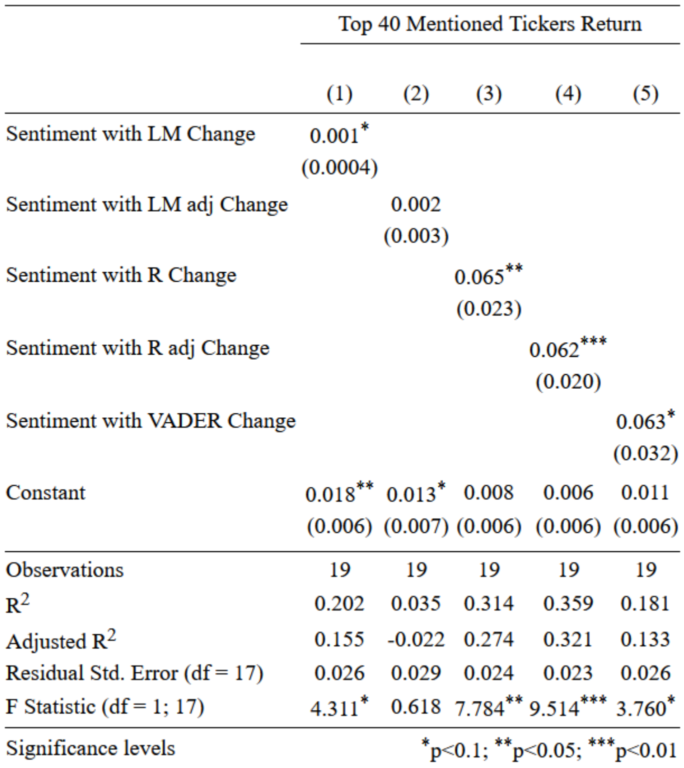
7.4 - Comparison between different word lists

To check whether the values derived from the use of R actually are better adapted to the dependent variable, a simple linear regression is carried out. The dependent variable concerns the returns of stocks with more than 500 mentions while the independent variables concern the variation in the results of the sentiment analysis deriving from the use of the different word lists employed in this research from 16th May 2021 to 11th June 2021 (Figure 7.10 based on Table 10). First of all, F Statistic turns out statistically significant for the fourth model because the p-value is lower than 0.01 and for the third because it is lower than 0.05. While the second model is rejected a priori, the first and fifth are not very significant (having a p-value higher than 0.05). Therefore, you might not consider them for the purposes of the analysis. Since the two chosen models (models 3 and 4) have only one independent variable, I consider the R^2 which is moderate but substantially good given that the models are based on a sentiment analysis which is not quantitative but qualitative and, as a consequence, not perfect as well as the small number of observations available. From the two models it can be stated that:

- Model (3): an increase of one percentage point in the redditors' sentiment using the R word list produces an average increase of 0.065 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.

- Model (4): an increase of one percentage point in the redditors' sentiment using the adjusted R word list produces an average increase of 0.062 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.

Figure 7.10: Linear regressions between stocks return and sentiment change.

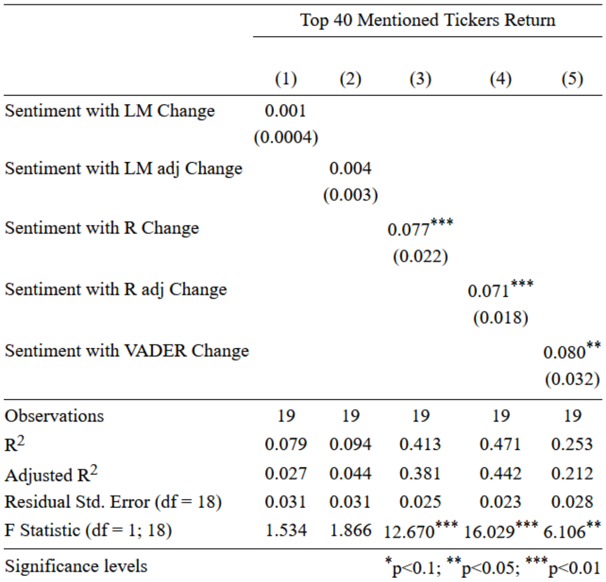


	Top 40 Mentioned Tickers Return				
	(1)	(2)	(3)	(4)	(5)
Sentiment with LM Change	0.001*				
	(0.0004)				
Sentiment with LM adj Change		0.002			
		(0.003)			
Sentiment with R Change			0.065**		
			(0.023)		
Sentiment with R adj Change				0.062***	
				(0.020)	
Sentiment with VADER Change					0.063*
					(0.032)
Constant	0.018**	0.013*	0.008	0.006	0.011
	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)
Observations	19	19	19	19	19
R^2	0.202	0.035	0.314	0.359	0.181
Adjusted R^2	0.155	-0.022	0.274	0.321	0.133
Residual Std. Error (df = 17)	0.026	0.029	0.024	0.023	0.026
F Statistic (df = 1; 17)	4.311*	0.618	7.784**	9.514***	3.760*
Significance levels	*p<0.1; **p<0.05; ***p<0.01				

Given that the constant in all the models which have been taken into consideration (models 3 and 4) is not statistically significant, you can therefore proceed in the evaluation of the regression models without constants (Figure 7.11 based on Table 10). First of all, F Statistic turns out statistically significant for the third and fourth models because the p-value is lower than 0.01 and the fifth model because the p-value is lower than 0.05. It is important to highlight that R^2 can also be evaluated in models without an intercept/constant. However, the interpretation of this parameter cannot be compared between models with and without intercept/constant. Therefore, only as regards the models without constant/intercept, once again they seem to have a moderate R^2. Then proceeding in the evaluation of the individual parameters, it can be inferred that these are all statistically significant. From the three models it can be stated that:

- Model (3): an increase of one percentage point in the redditors’ sentiment using the R word list produces an average increase of 0.077 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.
- Model (4): an increase of one percentage point in the redditors’ sentiment using the adjusted R word list produces an average increase of 0.071 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.
- Model (5): an increase of one percentage point in the redditors’ sentiment using the VADER produces an average increase of 0.080 percentage points in the 40 most mentioned companies return, therefore it has a positive effect.

Figure 7.11: Linear regressions without constants between stocks return and sentiment change.



	Top 40 Mentioned Tickers Return				
	(1)	(2)	(3)	(4)	(5)
Sentiment with LM Change	0.001				
	(0.0004)				
Sentiment with LM adj Change		0.004			
		(0.003)			
Sentiment with R Change			0.077^{***}		
			(0.022)		
Sentiment with R adj Change				0.071^{***}	
				(0.018)	
Sentiment with VADER Change					0.080^{**}
					(0.032)
Observations	19	19	19	19	19
R^2	0.079	0.094	0.413	0.471	0.253
Adjusted R^2	0.027	0.044	0.381	0.442	0.212
Residual Std. Error (df = 18)	0.031	0.031	0.025	0.023	0.028
F Statistic (df = 1; 18)	1.534	1.866	12.670^{***}	16.029^{***}	6.106^{**}
Significance levels			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

CONCLUSION

Even if the figures regarding unstructured data might be apparently insurmountable and discouraging, exploiting the potential of such data has now become an essential challenge in all sectors because it offers new opportunities for innovation, growth and competitive advantage. In the financial sector, the unstructured data analysed to extract the sentiment of market participants and commentators mainly comes from three types of sources: public corporate disclosures/filings, media articles and internet messages. The sentiment of investors expressed in comments on social media captures their subjective opinions and their possible consequent behaviours, which you need to quantify in order to check the effects on the trend and performance of the individual stocks and on the overall market. The short-term effects of this qualitative data on the market variables such as stock prices, returns, trading volumes and volatility can provide new complementary and incremental information to classic quantitative data (Loughran and McDonald 2016). In this research, general financial discussions such as "Daily Discussion Thread", "What Are Your Moves Tomorrow", "Weekend Discussion Thread" contained in the WSB subreddit were analysed. The initial part of the empirical section shows a certain weekly seasonality as the majority of comments are concentrated when the Stock Exchange is open while a very few of them are concentrated during the weekends. Whenever the stock market is open redditors write brief comments of the same length while, on the contrary, they write longer and more substantial comments during the weekends. In spite of the generality of the posts, the use of n-grams and post-tagging made it possible to collect the most discussed topics in the comments of the dataset in the period of time taken into consideration. Both methods have put in evidence and confirmed that the interest of the majority of users mainly converge towards certain listed companies mentioning the relevant tickers, such as AMC Entertainment Holdings Inc (AMC), BlackBerry (BB), GameStop (GME), Clover Health Investments Corp (CLOV), Clean Energy Fuels Corp (CLNE). It is notable that the more mentions increase the more share price does the same while, whenever mentions start to decrease also the price tends to decrease. The research of the tickers of companies listed on the NYSE and NASDAQ and the corresponding daily comments count of mentions made it possible to meet the challenge shown as the research of the future by Cruz, Kinyua and Mutigwe (2023), namely to extend the analysis to a wider range of companies. Using the most common way used by researchers through linear regressions (Kearney and Liu 2014) to explain the dependent variable of firm-level or market-level performance measure it is found that the more the variation in mentions, sentiment and group identity increase the more the effect on the 40 most mentioned companies return is positive. These results confirm that those companies that the media and business newspapers point out as the most influenced by social media in their stock performance correspond to those most cited also in the comments of redditors. Once again, the more the positive sentiment concerning these companies increases, the more the shares and their price increase, which shows a greater aggregation and cohesion among redditors. The verification of group identity shows the presence of a strong sense of belonging to the group of users, which can direct their interests and actions also on a collective level. The extraction of sentiment from users' comments confirmed the incapability of the LM word list for accurately catching their opinion. The use of LM confirms what the authors themselves state, namely that they are not able to accurately capture the sentiment contained in dynamic comments on social media (Loughran and McDonald 2016). Using R – specifically developed on the language of social media – allows one to obtain more precise results (Renault 2017). The sentiment resulting from R shows a pattern similar to the stock returns of the listed companies mostly mentioned by redditors. Adding the most used emojis and jargon in WSB enables us to extend the sentiment analysis (Novak et al., 2017) given that 10.92% of the comments in the total dataset contain at least one emoji. There are approximately four times as many emojis typically used in WSB that reflect a positive sentiment compared to those that express a negative sentiment. Their usage is more evident in conjunction with peaks relating to the mentions of the most discussed listed companies, changes in the prices of the related shares and group identity.

Unfortunately, this kind of research is limited by the fact that the APIs of some social media used to extract comments have recently undergone an abrupt change in policies with a consequent high increase in usage costs. This change, which aims at opposing AI training, has made it more complicated to carry out textual analysis of unstructured data (Isaac 2023).

For a future research, content analysis methods could be developed and improved by expert linguists, psychologists and computer scientists to measure sentiment from unstructured data in a more accurate way, maybe in languages other than English. Furthermore, sentiment analysis related to the stock market could be extended to other types of markets such as bonds and derivatives. In addition, public corporate releases occur with a low frequency on an annual or quarterly basis, while media articles and internet messages are much more frequent. Therefore, to evaluate the effect of sentiment on stock returns it may be appropriate to use the three different sources of information jointly. This way, one can obtain more granular analysis at a weekly, daily and even intraday level. Each type of information source requires the use of a word list appropriate to the context. The results thus obtained must then be grouped into a single measurement.

APPENDIX A - Popular WallStreetBets emojis

Emojis with positive meaning		
Emoji Name	**Emoji**	**Meaning**
Rocket	🚀	Enthusiasm for the performance of a stock that has earned money or is rising sharply; it is often accompanied by hands
Moon	🌕 🌑 🌚 🌙	As above
Chart Increasing	📈	As above
Poultry Leg	🍗	Joking. Chicken fingers to eat to celebrate the profits obtained from an investment
Money Bag	💰	Cash in a large profit
Dollar Banknote	💵	As above
Money-Mouth Face	🤑	Very profitable investment
Bullseye	🎯	Hit the target by making a good investment
Ox	🐂	Those who think positively about a certain investment
Gem Stone	💎	Positive attitude towards a determined stock, that it is intended to maintain in all cases
Gorilla / Monkey	🦍 🐒	When the monkey is in a group it inspires fear
Flexed Biceps	💪	Strength, resistance, skill, power
Fire	🔥	Strong interest in a stock that is doing well
Star-Struck	🤩	Mind-blowing, surprising, astounding
Thumbs Up	👍	Approval

Emojis with negative meaning		
Emoji Name	**Emoji**	**Meaning**
Chart Decreasing	📉	It represents an investment that is decreasing in value
Bear / Teddy Bear	🐻 🧸	It refers to a bearish investor who, by selling, causes the value of an investment to fall even further
Roll of Paper	🧻	Those who sell their shares at the slightest drop; it is accompanied by the hands
Whale / Spouting Whale	🐋 🐳	Investor who owns so many shares that, if he/she sells them, he/she can cause their price to drop
Loudly Crying Face	😭	Crying for bad investments
Face Screaming in Fear	😱	Scream of fear, terror
Face with Steam from Nose	😤	Irritation, impatience, frustration
Angry Face	😠	Anger, disgust, indignation
Enraged Face	😡	Alteration, anger, fury
Clown Face	🤡	Insult or dissent towards something or someone
Skull / Skull and Crossbones	💀 ☠️	Danger, ruin, desperation, end
Thumbs Down	👎	Disapproval

APPENDIX B - Popular WallStreetBets jargon – (p) positive, (n) negative, (i) indefinite

$Becky – It indicates a set of stocks that have a strong demand among young white women, mostly university students. (p)

$ROPE – When an investor has lost a sizeable significant amount, he is sarcastically asked to buy a rope. (n)

2 bagger – Investment that doubles the initial amount. (p)

5 bagger / **five-bagger** / **five bagger** – Investment that has quintupled the initial amount. (p)

10 bagger – Investment with a return equal to 1000%, i.e. 10 times the initial amount. The term comes from the world of baseball. (p)

Andromeda – One of the terms used to express the belief that a specific particular stock will have a significant increase. (p)

Apes / **Apes together strong** – Phrase taken from a meme referring to the film *Rise of the Planet of the Apes* of 2011. The WallStreetBets community identifies with the monkey, an animal that alone does not inspire fear but in a group becomes a force. (p)

ATH – acronym for **All-Time High**. It refers to a stock or cryptocurrency that has reached its new maximum. (p)

ATL – acronym for **All-Time Low**. It refers to a stock or cryptocurrency that has reached its new minimum. (n)

Autism / Autistic / Autist – Those who work with advanced techniques and consistently profit from their analyses. (p)

Bagholder / Bag holder / Bag-holder – An investor who owns consistently losing stocks. (n)

Ban – Request made to the moderator to ban the author of a statement deemed enormously senseless. (n)

BANG – Acronym that groups together four stocks much discussed in the WallStreetBets community: BlackBerry (BB), AMC Entertainment (AMC), Nokia (NOK) e GameStop (GME). (i)

Bear / Bears – Bearish investor. Convinced that a stock will decrease in value he/she tends to sell causing it to drop even further. This is how they named it because in the wild the bear attacks with a downward movement. (n)

Bear gang – It indicates who is always bearish. (n)

Bearish – The term defines someone who is convinced that a stock price will fall. Those who are always bearish are part of the **Bear gang** (see). (n)

Black swan / Blacksvan – An absolutely unthinkable event, but evident with the benefit of hindsight, from which even serious consequences can arise. (n)

BTFD – Acronym which invites you to buy shares that have fallen but are expected to rise. See **Buy the dip**. (p)

Bull gang – Those who are part of the **Bull gang** are always bullish. (p)

Bullish – Who thinks that a given stock or cryptocurrency is profitable. Those who are always bullish are part of the **Bull gang** (see). (p)

Buy high sell low – Joking phrase which is said when an investor shares the extent of his/her losses on the forum. They bought high and sold low. (n)

Buy the dip – Expression which invites you to buy a stock immediately after a temporary drop in price, in the hope that it will soon rise again. See **BTFD**. (p)

Copium – This term indicates comments that are contrary to or do not agree with market movements. (n)

DD – Shortening of term **Due Diligence**. It is used by the redditors in order to indicate that before investing they have been carried out searches on a determined stock or the market trend. Also used to mock a move deemed incorrect: "Great DD on that one…". (p)

Diamond hands – It refers to investors who intend to hold a stock for a long time in the belief that sooner or later it will increase. No pressure will bend them and their strength is equal to that of the diamond. It was depicted by diamond and hands emojis. Opposite of **Paper hands** (see). (p)

Drilling / Drill team 6 – They indicate a stock that has fallen either by a little or by a lot. (n)

FD / **Fds** – Offensive acronyms that refer to investments that have weekly maturity and are very risky but with high probability of profit. (i)

FOMO / **Fear Of Missing Out** – Acronym that indicates the fear of losing. (n)

Gain pn** – Screenshot sharing the accounts of a trade that resulted in a sizeable and meaningful gain, which must be $10,000 or more. (p)

G*y bears – Offensive term used to refer to bearish investors. (n)

HODL / Hodling – From a typo for "hold". It refers to the strategy of maintaining an investment for a long time, without being intimidated by short-term market fluctuations, in the belief that there will be a profit. (p)

Hold the line / Holding the line – Invitation to hold a stock when it is decreasing in value. (p)

In scrambles – It refers to someone who has made a wrong decision and regretted it. (n)

It’s not a loss until you sell – Even if the value of an investment has dropped, as long as you don't sell it is still not a loss. There is always the hope that the value will rise. (i)

JPOW / Daddy – Nickname used to refer to **Jerome Powell**, the Chairman of the Federal Reserve, seen positively because he had lowered interest rates. (i)

Long as a python – Optimistic statement referring to a prolonged or intense bullish solid trend in a single company, often speculative, without considering the risk. (p)

Loss pn** – Screenshot sharing the accounts of a trade that resulted in a sizeable and meaningful loss, which must be $10,000 or more. (n)

Meme stock – It is a stock that has gone viral on forums or social media and, precisely because of this attention, sees its prices rise. (p)

Moon / Mooning – It indicates a stock that has risen, slightly or significantly, until it reaches the moon. (p)

Multibagger – It refers to investments in shares of companies with great development potential which have a return several times higher than their cost. (p)

Paper hands – It refers to investors who sell at the first sign of the downtrend. Derogatory towards those who do not maintain their actions. Represented by a roll of paper and hands. Opposite of **Diamond hands** (see). (n)

Retard – Offensive term for telling someone who doesn't know what they are doing. (n)

Rocket / Rockets / Rocket ships – With these words, which can also be accompanied by the relevant emoji, the belief in the strong growth of a specific stock is expressed. (p)

Short squeeze – It occurs when a security, sold largely short, undergoes a rapid and unexpected surge in price, which forces a further purchase, resulting in a further increase in value. (p)

Smooth brain – Offensive term. (n)

Stonks / Not Stonks – Intentional misspelling of “stocks”. This term refers to stocks and the financial world in an ironic way, especially with regards to monetary gains or losses. (p) / (n)

Tendies – Jokingly referring to chicken fingers, which investors intend to eat to celebrate gains made on an investment. The term is depicted by the emoji of a chicken leg. (p)

To the moon – This phrase expresses enthusiasm for the performance of a stock that has gained money or the prediction of a solid substantial increase in its value. It is usually accompanied by a rocket emoji and sometimes even a full moon emoji. (p)

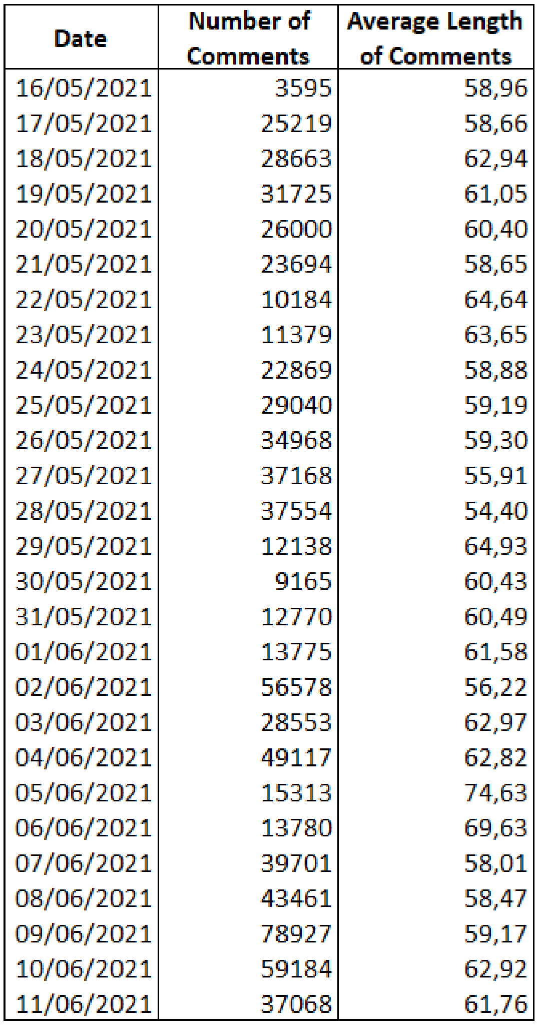
We like the stock / I like the stock / I just like the stock – The phrase, initially uttered by a television presenter and sometimes repeated in sequence, refers positively to a trending stock. (p)

Whale – An investor who owns so many stocks that if he/she sells them, he/she can cause their price to drop. (n)

YOLO – acronym for **You Only Live Once**. Expression used when investing a large sum in a single stock. Impulsive, perilous decision similar to gambling. (p)

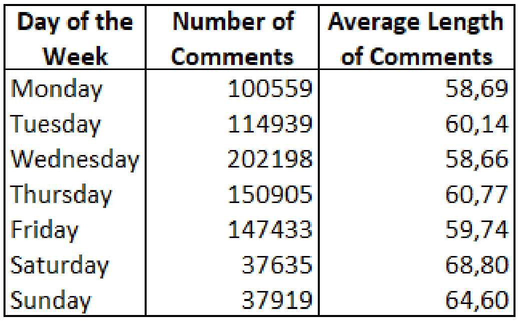
TABLES

Table 1



Date	Number of Comments	Average Length of Comments
16/05/2021	3595	58,96
17/05/2021	25219	58,66
18/05/2021	28663	62,94
19/05/2021	31725	61,05
20/05/2021	26000	60,40
21/05/2021	23694	58,65
22/05/2021	10184	64,64
23/05/2021	11379	63,65
24/05/2021	22869	58,88
25/05/2021	29040	59,19
26/05/2021	34968	59,30
27/05/2021	37168	55,91
28/05/2021	37554	54,40
29/05/2021	12138	64,93
30/05/2021	9165	60,43
31/05/2021	12770	60,49
01/06/2021	13775	61,58
02/06/2021	56578	56,22
03/06/2021	28553	62,97
04/06/2021	49117	62,82
05/06/2021	15313	74,63
06/06/2021	13780	69,63
07/06/2021	39701	58,01
08/06/2021	43461	58,47
09/06/2021	78927	59,17
10/06/2021	59184	62,92
11/06/2021	37068	61,76

Table 2



Day of the Week	Number of Comments	Average Length of Comments
Monday	100559	58,69
Tuesday	114939	60,14
Wednesday	202198	58,66
Thursday	150905	60,77
Friday	147433	59,74
Saturday	37635	68,80
Sunday	37919	64,60

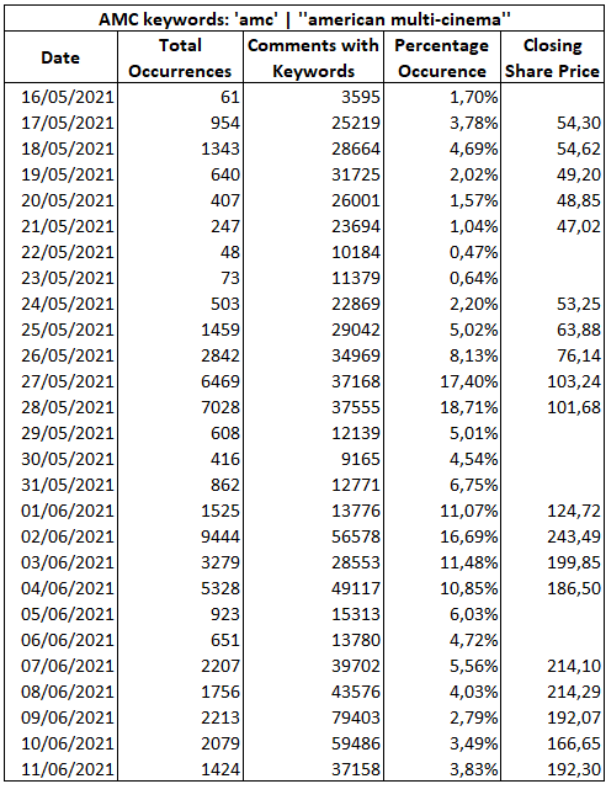
Table 3

Positive emojis		Negative emojis	
Emoji (name)	**Occurences**	**Emoji (name)**	**Occurences**
🚀 (Rocket)	25698	🐻 (Bear)	4183
🦍 (Gorilla)	4815	🤡 (Clown Face)	4151
💎 (Gem Stone)	4344	😭 (Loudly Crying Face)	950
🔥 (Fire)	2337	💀 (Skull)	502
🤑 (Money-Mouth Face)	1331	😤 (Face with Steam from Nose)	475
🌜 (Crescent Moon)	1093	📉 (Chart Decreasing)	235
📈 (Chart Increasing)	982	🐋 (Whale)	195
👍 (Thumbs Up)	892	😡 (Enraged Face)	188
🌕 (Full Moon)	774	😱 (Face Screaming in Fear)	146
💰 (Money Bag)	771	☠ (Skull and Crossbones)	86
💪 (Flexed Biceps)	574	🧻 (Roll of Paper)	79
🌚 (New Moon Face)	508	🧸 (Teddy Bear)	63
🐒 (Monkey)	294	😠 (Angry Face)	54
💵 (Dollar Banknote)	240	🐳 (Spouting Whale)	46
🐂 (Ox)	238	👎 (Thumbs Down)	42
🤩 (Star-Struck)	191		
🌑 (New Moon)	104		
🎯 (Bullseye)	72		
🍗 (Poultry Leg)	59		
Total	45317	**Total**	11395

Table 4

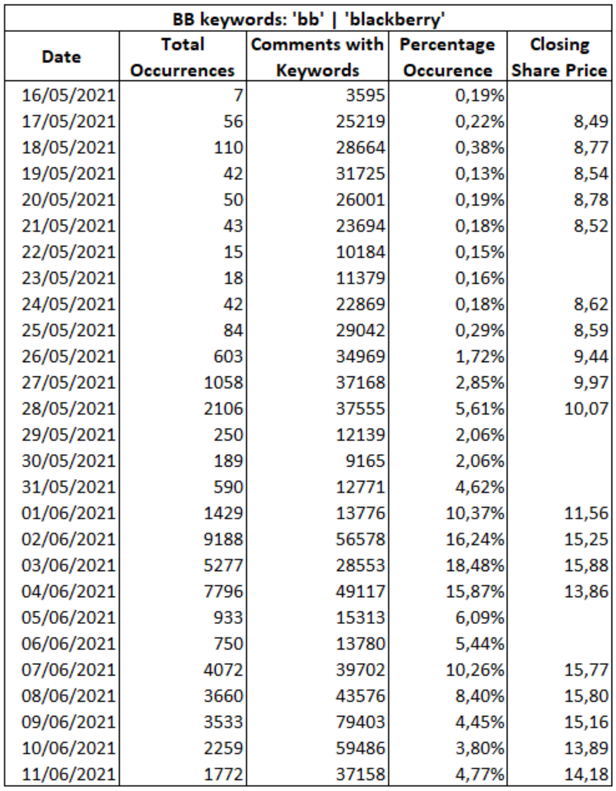
		amc		bb		gme		clov		clne			
Date	**Number of Comments**	**Number of Ticker Mentions**	**Mention Percentage**	**Number of Ticker Mentions**	**Mention Percentage**	**Number of Ticker Mentions**	**Mention Percentage**	**Number of Ticker Mentions**	**Mention Percentage**	**Number of Ticker Mentions**	**Mention Percentage**	**Average Top 5 Most Mentioned Tickers**	**Group Identity**
16/05/2021	3595	0	0,00%	0	0,00%	0	0,00%	0	0,00%	0	0,00%		13,24%
17/05/2021	25219	909	3,60%	0	0,00%	1218	4,83%	0	0,00%	0	0,00%	4,22%	13,69%
18/05/2021	28664	1423	4,96%	0	0,00%	1522	5,31%	0	0,00%	0	0,00%	5,14%	14,41%
19/05/2021	31725	614	1,94%	0	0,00%	633	2,00%	0	0,00%	0	0,00%	1,97%	16,05%
20/05/2021	26001	0	0,00%	0	0,00%	0	0,00%	0	0,00%	0	0,00%		14,29%
21/05/2021	23694	0	0,00%	0	0,00%	0	0,00%	0	0,00%	0	0,00%		13,75%
22/05/2021	10184	0	0,00%	0	0,00%	0	0,00%	0	0,00%	0	0,00%		12,02%
23/05/2021	11379	0	0,00%	0	0,00%	0	0,00%	0	0,00%	0	0,00%		14,58%
24/05/2021	22869	471	2,06%	0	0,00%	0	0,00%	0	0,00%	0	0,00%	2,06%	14,31%
25/05/2021	29042	1428	4,92%	0	0,00%	2226	7,66%	0	0,00%	0	0,00%	6,29%	16,06%
26/05/2021	34969	2660	7,61%	0	0,00%	3124	8,93%	0	0,00%	0	0,00%	8,27%	15,93%
27/05/2021	37168	6202	16,69%	1053	2,83%	2511	6,76%	0	0,00%	0	0,00%	8,76%	17,00%
28/05/2021	37555	6797	18,10%	2051	5,46%	1510	4,02%	0	0,00%	0	0,00%	9,19%	19,20%
29/05/2021	12139	568	4,68%	234	1,93%	0	0,00%	0	0,00%	0	0,00%	3,30%	11,49%
30/05/2021	9165	398	4,34%	180	1,96%	0	0,00%	0	0,00%	0	0,00%	3,15%	13,48%
31/05/2021	12771	814	6,37%	540	4,23%	239	1,87%	0	0,00%	0	0,00%	4,16%	15,09%
01/06/2021	13776	1486	10,79%	1434	10,41%	409	2,97%	0	0,00%	0	0,00%	8,06%	14,47%
02/06/2021	56578	9112	16,11%	9374	16,57%	2518	4,45%	0	0,00%	0	0,00%	12,37%	18,33%
03/06/2021	28553	3243	11,36%	5298	18,55%	988	3,46%	0	0,00%	0	0,00%	11,12%	16,73%
04/06/2021	49117	5310	10,81%	7920	16,12%	1283	2,61%	0	0,00%	1223	2,49%	8,01%	18,03%
05/06/2021	15313	945	6,17%	1020	6,66%	350	2,29%	0	0,00%	0	0,00%	5,04%	14,45%
06/06/2021	13780	652	4,73%	774	5,62%	307	2,23%	0	0,00%	412	2,99%	3,89%	13,80%
07/06/2021	39702	2136	5,38%	4088	10,30%	1291	3,25%	2189	5,51%	1688	4,25%	5,74%	17,39%
08/06/2021	43576	1692	3,88%	3664	8,41%	1822	4,18%	6556	15,04%	1339	3,07%	6,92%	18,91%
09/06/2021	79403	2174	2,74%	3565	4,49%	0	0,00%	5736	7,22%	5280	6,65%	5,28%	19,15%
10/06/2021	59486	2006	3,37%	2356	3,96%	0	0,00%	2282	3,84%	2579	4,34%	3,88%	18,29%
11/06/2021	37158	1407	3,79%	1854	4,99%	0	0,00%	1915	5,15%	1178	3,17%	4,27%	17,82%

Table 5



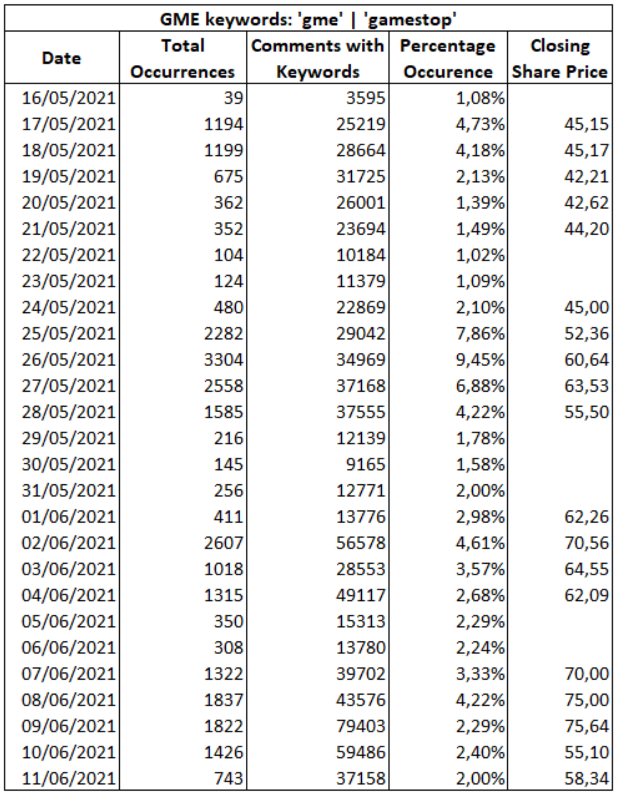
AMC keywords: 'amc' | ''american multi-cinema''				
Date	Total	Comments with	Percentage	Closing
	Occurrences	Keywords	Occurence	Share Price
16/05/2021	61	3595	1,70%	
17/05/2021	954	25219	3,78%	54,30
18/05/2021	1343	28664	4,69%	54,62
19/05/2021	640	31725	2,02%	49,20
20/05/2021	407	26001	1,57%	48,85
21/05/2021	247	23694	1,04%	47,02
22/05/2021	48	10184	0,47%	
23/05/2021	73	11379	0,64%	
24/05/2021	503	22869	2,20%	53,25
25/05/2021	1459	29042	5,02%	63,88
26/05/2021	2842	34969	8,13%	76,14
27/05/2021	6469	37168	17,40%	103,24
28/05/2021	7028	37555	18,71%	101,68
29/05/2021	608	12139	5,01%	
30/05/2021	416	9165	4,54%	
31/05/2021	862	12771	6,75%	
01/06/2021	1525	13776	11,07%	124,72
02/06/2021	9444	56578	16,69%	243,49
03/06/2021	3279	28553	11,48%	199,85
04/06/2021	5328	49117	10,85%	186,50
05/06/2021	923	15313	6,03%	
06/06/2021	651	13780	4,72%	
07/06/2021	2207	39702	5,56%	214,10
08/06/2021	1756	43576	4,03%	214,29
09/06/2021	2213	79403	2,79%	192,07
10/06/2021	2079	59486	3,49%	166,65
11/06/2021	1424	37158	3,83%	192,30

Table 6



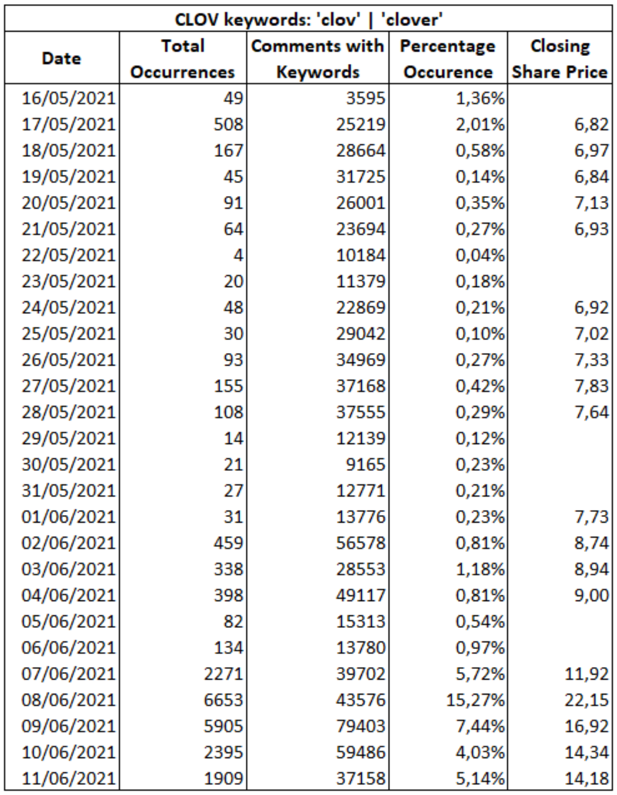
BB keywords: 'bb' | 'blackberry'				
Date	Total	Comments with	Percentage	Closing
	Occurrences	Keywords	Occurence	Share Price
16/05/2021	7	3595	0,19%	
17/05/2021	56	25219	0,22%	8,49
18/05/2021	110	28664	0,38%	8,77
19/05/2021	42	31725	0,13%	8,54
20/05/2021	50	26001	0,19%	8,78
21/05/2021	43	23694	0,18%	8,52
22/05/2021	15	10184	0,15%	
23/05/2021	18	11379	0,16%	
24/05/2021	42	22869	0,18%	8,62
25/05/2021	84	29042	0,29%	8,59
26/05/2021	603	34969	1,72%	9,44
27/05/2021	1058	37168	2,85%	9,97
28/05/2021	2106	37555	5,61%	10,07
29/05/2021	250	12139	2,06%	
30/05/2021	189	9165	2,06%	
31/05/2021	590	12771	4,62%	
01/06/2021	1429	13776	10,37%	11,56
02/06/2021	9188	56578	16,24%	15,25
03/06/2021	5277	28553	18,48%	15,88
04/06/2021	7796	49117	15,87%	13,86
05/06/2021	933	15313	6,09%	
06/06/2021	750	13780	5,44%	
07/06/2021	4072	39702	10,26%	15,77
08/06/2021	3660	43576	8,40%	15,80
09/06/2021	3533	79403	4,45%	15,16
10/06/2021	2259	59486	3,80%	13,89
11/06/2021	1772	37158	4,77%	14,18

Table 7



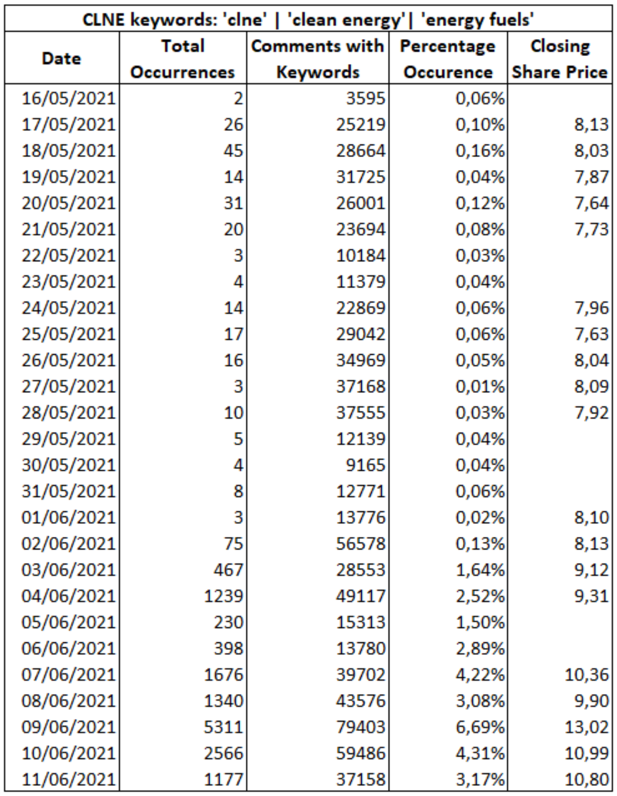
GME keywords: 'gme' | 'gamestop'				
Date	Total	Comments with	Percentage	Closing
	Occurrences	Keywords	Occurence	Share Price
16/05/2021	39	3595	1,08%	
17/05/2021	1194	25219	4,73%	45,15
18/05/2021	1199	28664	4,18%	45,17
19/05/2021	675	31725	2,13%	42,21
20/05/2021	362	26001	1,39%	42,62
21/05/2021	352	23694	1,49%	44,20
22/05/2021	104	10184	1,02%	
23/05/2021	124	11379	1,09%	
24/05/2021	480	22869	2,10%	45,00
25/05/2021	2282	29042	7,86%	52,36
26/05/2021	3304	34969	9,45%	60,64
27/05/2021	2558	37168	6,88%	63,53
28/05/2021	1585	37555	4,22%	55,50
29/05/2021	216	12139	1,78%	
30/05/2021	145	9165	1,58%	
31/05/2021	256	12771	2,00%	
01/06/2021	411	13776	2,98%	62,26
02/06/2021	2607	56578	4,61%	70,56
03/06/2021	1018	28553	3,57%	64,55
04/06/2021	1315	49117	2,68%	62,09
05/06/2021	350	15313	2,29%	
06/06/2021	308	13780	2,24%	
07/06/2021	1322	39702	3,33%	70,00
08/06/2021	1837	43576	4,22%	75,00
09/06/2021	1822	79403	2,29%	75,64
10/06/2021	1426	59486	2,40%	55,10
11/06/2021	743	37158	2,00%	58,34

Table 8



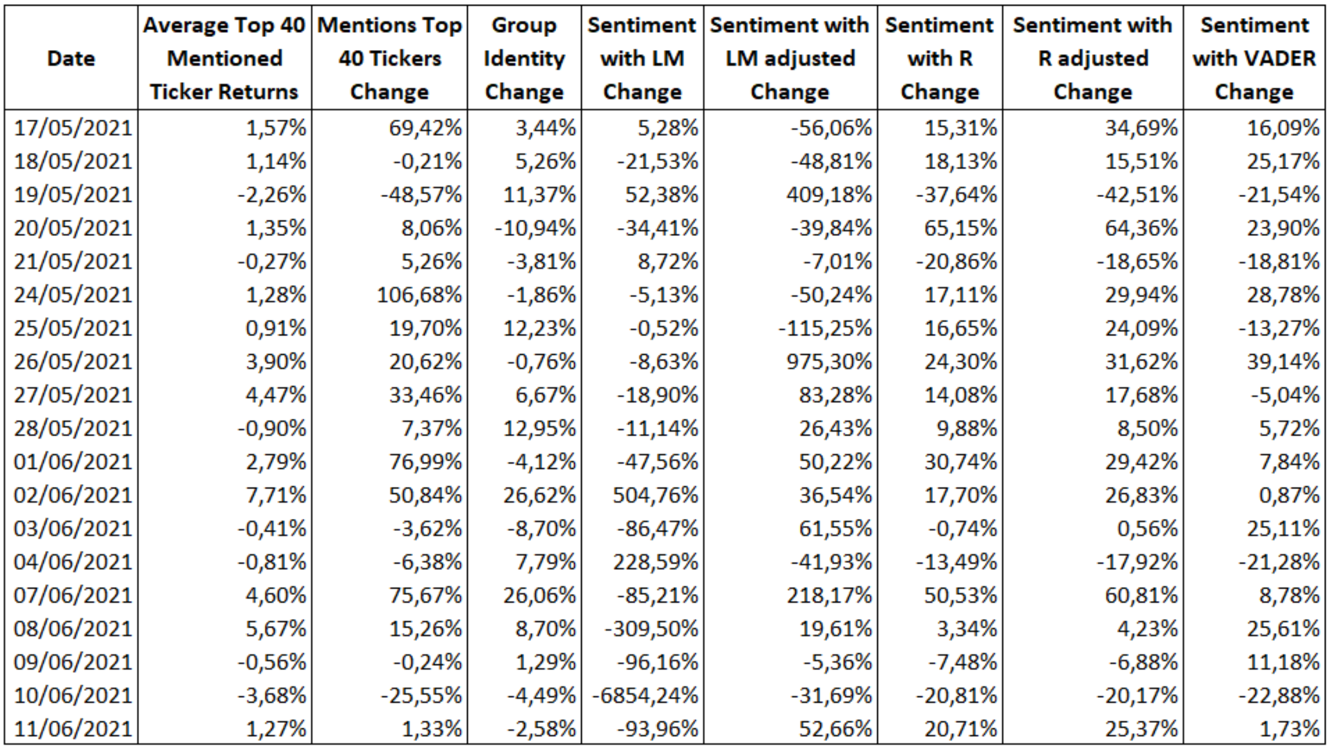
CLOV keywords: 'clov' | 'clover'				
Date	Total	Comments with	Percentage	Closing
	Occurrences	Keywords	Occurence	Share Price
16/05/2021	49	3595	1,36%	
17/05/2021	508	25219	2,01%	6,82
18/05/2021	167	28664	0,58%	6,97
19/05/2021	45	31725	0,14%	6,84
20/05/2021	91	26001	0,35%	7,13
21/05/2021	64	23694	0,27%	6,93
22/05/2021	4	10184	0,04%	
23/05/2021	20	11379	0,18%	
24/05/2021	48	22869	0,21%	6,92
25/05/2021	30	29042	0,10%	7,02
26/05/2021	93	34969	0,27%	7,33
27/05/2021	155	37168	0,42%	7,83
28/05/2021	108	37555	0,29%	7,64
29/05/2021	14	12139	0,12%	
30/05/2021	21	9165	0,23%	
31/05/2021	27	12771	0,21%	
01/06/2021	31	13776	0,23%	7,73
02/06/2021	459	56578	0,81%	8,74
03/06/2021	338	28553	1,18%	8,94
04/06/2021	398	49117	0,81%	9,00
05/06/2021	82	15313	0,54%	
06/06/2021	134	13780	0,97%	
07/06/2021	2271	39702	5,72%	11,92
08/06/2021	6653	43576	15,27%	22,15
09/06/2021	5905	79403	7,44%	16,92
10/06/2021	2395	59486	4,03%	14,34
11/06/2021	1909	37158	5,14%	14,18

Table 9



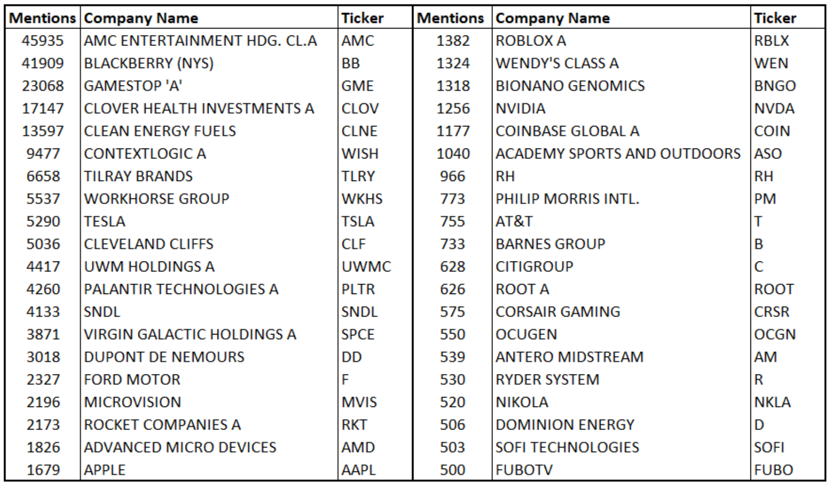
CLNE keywords: 'clne' | 'clean energy'| 'energy fuels'				
Date	Total Occurrences	Comments with Keywords	Percentage Occurence	Closing Share Price
16/05/2021	2	3595	0,06%	
17/05/2021	26	25219	0,10%	8,13
18/05/2021	45	28664	0,16%	8,03
19/05/2021	14	31725	0,04%	7,87
20/05/2021	31	26001	0,12%	7,64
21/05/2021	20	23694	0,08%	7,73
22/05/2021	3	10184	0,03%	
23/05/2021	4	11379	0,04%	
24/05/2021	14	22869	0,06%	7,96
25/05/2021	17	29042	0,06%	7,63
26/05/2021	16	34969	0,05%	8,04
27/05/2021	3	37168	0,01%	8,09
28/05/2021	10	37555	0,03%	7,92
29/05/2021	5	12139	0,04%	
30/05/2021	4	9165	0,04%	
31/05/2021	8	12771	0,06%	
01/06/2021	3	13776	0,02%	8,10
02/06/2021	75	56578	0,13%	8,13
03/06/2021	467	28553	1,64%	9,12
04/06/2021	1239	49117	2,52%	9,31
05/06/2021	230	15313	1,50%	
06/06/2021	398	13780	2,89%	
07/06/2021	1676	39702	4,22%	10,36
08/06/2021	1340	43576	3,08%	9,90
09/06/2021	5311	79403	6,69%	13,02
10/06/2021	2566	59486	4,31%	10,99
11/06/2021	1177	37158	3,17%	10,80

Table 10



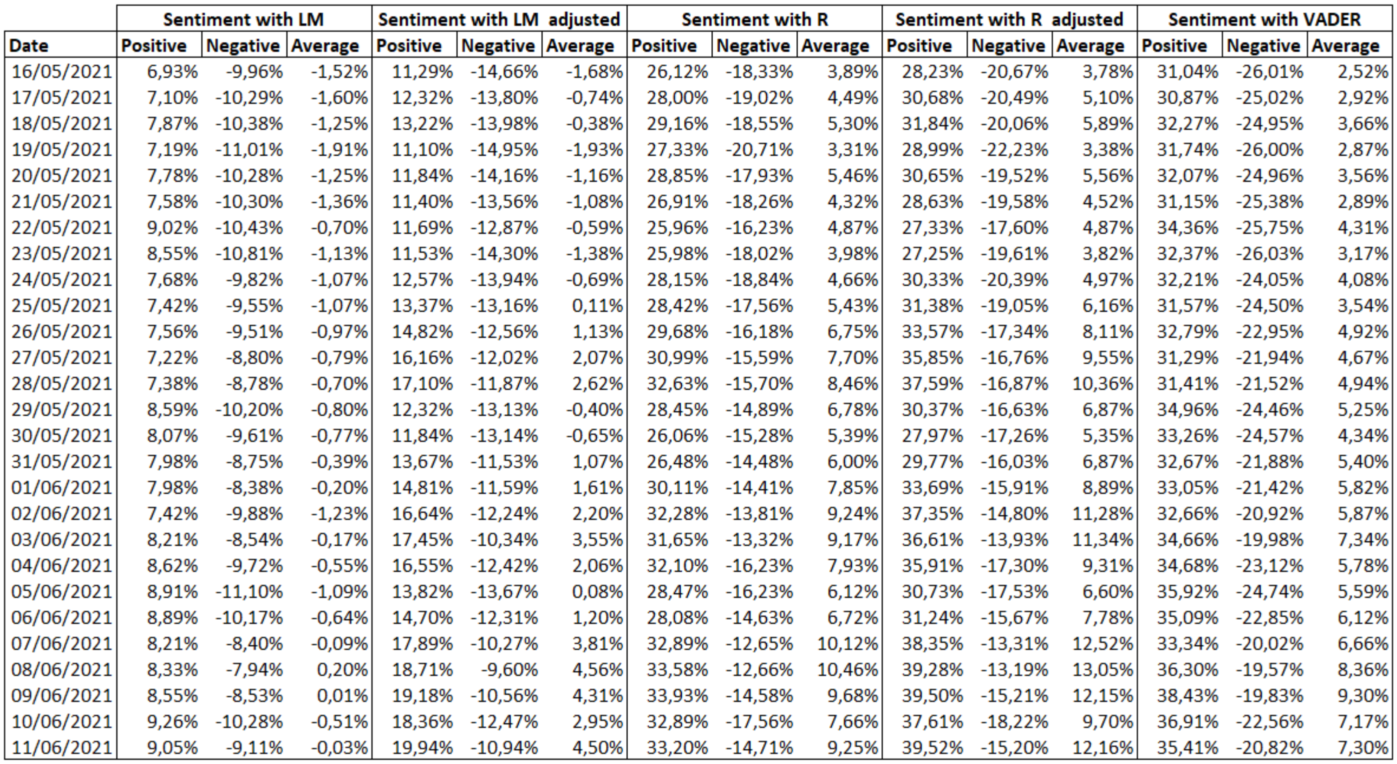
Date	Average Top 40 Mentioned Ticker Returns	Mentions Top 40 Tickers Change	Group Identity Change	Sentiment with LM Change	Sentiment with LM adjusted Change	Sentiment with R Change	Sentiment with R adjusted Change	Sentiment with VADER Change
17/05/2021	1,57%	69,42%	3,44%	5,28%	-56,06%	15,31%	34,69%	16,09%
18/05/2021	1,14%	-0,21%	5,26%	-21,53%	-48,81%	18,13%	15,51%	25,17%
19/05/2021	-2,26%	-48,57%	11,37%	52,38%	409,18%	-37,64%	-42,51%	-21,54%
20/05/2021	1,35%	8,06%	-10,94%	-34,41%	-39,84%	65,15%	64,36%	23,90%
21/05/2021	-0,27%	5,26%	-3,81%	8,72%	-7,01%	-20,86%	-18,65%	-18,81%
24/05/2021	1,28%	106,68%	-1,86%	-5,13%	-50,24%	17,11%	29,94%	28,78%
25/05/2021	0,91%	19,70%	12,23%	-0,52%	-115,25%	16,65%	24,09%	-13,27%
26/05/2021	3,90%	20,62%	-0,76%	-8,63%	975,30%	24,30%	31,62%	39,14%
27/05/2021	4,47%	33,46%	6,67%	-18,90%	83,28%	14,08%	17,68%	-5,04%
28/05/2021	-0,90%	7,37%	12,95%	-11,14%	26,43%	9,88%	8,50%	5,72%
01/06/2021	2,79%	76,99%	-4,12%	-47,56%	50,22%	30,74%	29,42%	7,84%
02/06/2021	7,71%	50,84%	26,62%	504,76%	36,54%	17,70%	26,83%	0,87%
03/06/2021	-0,41%	-3,62%	-8,70%	-86,47%	61,55%	-0,74%	0,56%	25,11%
04/06/2021	-0,81%	-6,38%	7,79%	228,59%	-41,93%	-13,49%	-17,92%	-21,28%
07/06/2021	4,60%	75,67%	26,06%	-85,21%	218,17%	50,53%	60,81%	8,78%
08/06/2021	5,67%	15,26%	8,70%	-309,50%	19,61%	3,34%	4,23%	25,61%
09/06/2021	-0,56%	-0,24%	1,29%	-96,16%	-5,36%	-7,48%	-6,88%	11,18%
10/06/2021	-3,68%	-25,55%	-4,49%	-6854,24%	-31,69%	-20,81%	-20,17%	-22,88%
11/06/2021	1,27%	1,33%	-2,58%	-93,96%	52,66%	20,71%	25,37%	1,73%

Table 11



Mentions	Company Name	Ticker	Mentions	Company Name	Ticker
45935	AMC ENTERTAINMENT HDG. CL.A	AMC	1382	ROBLOX A	RBLX
41909	BLACKBERRY (NYS)	BB	1324	WENDY'S CLASS A	WEN
23068	GAMESTOP 'A'	GME	1318	BIONANO GENOMICS	BNGO
17147	CLOVER HEALTH INVESTMENTS A	CLOV	1256	NVIDIA	NVDA
13597	CLEAN ENERGY FUELS	CLNE	1177	COINBASE GLOBAL A	COIN
9477	CONTEXTLOGIC A	WISH	1040	ACADEMY SPORTS AND OUTDOORS	ASO
6658	TILRAY BRANDS	TLRY	966	RH	RH
5537	WORKHORSE GROUP	WKHS	773	PHILIP MORRIS INTL.	PM
5290	TESLA	TSLA	755	AT&T	T
5036	CLEVELAND CLIFFS	CLF	733	BARNES GROUP	B
4417	UWM HOLDINGS A	UWMC	628	CITIGROUP	C
4260	PALANTIR TECHNOLOGIES A	PLTR	626	ROOT A	ROOT
4133	SNDL	SNDL	575	CORSAIR GAMING	CRSR
3871	VIRGIN GALACTIC HOLDINGS A	SPCE	550	OCUGEN	OCGN
3018	DUPONT DE NEMOURS	DD	539	ANTERO MIDSTREAM	AM
2327	FORD MOTOR	F	530	RYDER SYSTEM	R
2196	MICROVISION	MVIS	520	NIKOLA	NKLA
2173	ROCKET COMPANIES A	RKT	506	DOMINION ENERGY	D
1826	ADVANCED MICRO DEVICES	AMD	503	SOFI TECHNOLOGIES	SOFI
1679	APPLE	AAPL	500	FUBOTV	FUBO

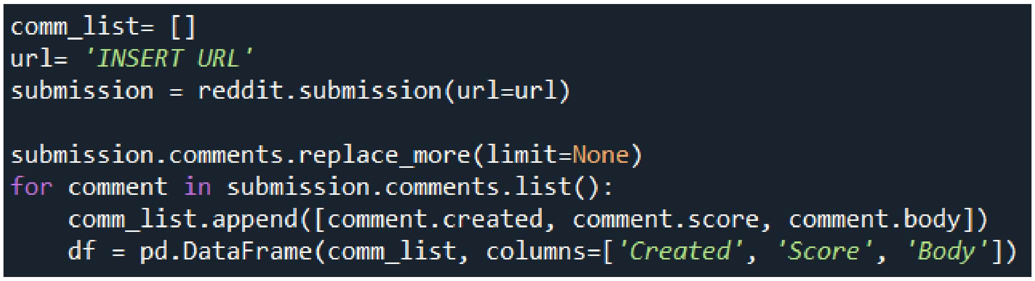
Table 12



	Sentiment with LM			Sentiment with LM adjusted			Sentiment with R			Sentiment with R adjusted			Sentiment with VADER		
Date	Positive	Negative	Average	Positive	Negative	Average	Positive	Negative	Average	Positive	Negative	Average	Positive	Negative	Average
16/05/2021	6,93%	-9,96%	-1,52%	11,29%	-14,66%	-1,68%	26,12%	-18,33%	3,89%	28,23%	-20,67%	3,78%	31,04%	-26,01%	2,52%
17/05/2021	7,10%	-10,29%	-1,60%	12,32%	-13,80%	-0,74%	28,00%	-19,02%	4,49%	30,68%	-20,49%	5,10%	30,87%	-25,02%	2,92%
18/05/2021	7,87%	-10,38%	-1,25%	13,22%	-13,98%	-0,38%	29,16%	-18,55%	5,30%	31,84%	-20,06%	5,89%	32,27%	-24,95%	3,66%
19/05/2021	7,19%	-11,01%	-1,91%	11,10%	-14,95%	-1,93%	27,33%	-20,71%	3,31%	28,99%	-22,23%	3,38%	31,74%	-26,00%	2,87%
20/05/2021	7,78%	-10,28%	-1,25%	11,84%	-14,16%	-1,16%	28,85%	-17,93%	5,46%	30,65%	-19,52%	5,56%	32,07%	-24,96%	3,56%
21/05/2021	7,58%	-10,30%	-1,36%	11,40%	-13,56%	-1,08%	26,91%	-18,26%	4,32%	28,63%	-19,58%	4,52%	31,15%	-25,38%	2,89%
22/05/2021	9,02%	-10,43%	-0,70%	11,69%	-12,87%	-0,59%	25,96%	-16,23%	4,87%	27,33%	-17,60%	4,87%	34,36%	-25,75%	4,31%
23/05/2021	8,55%	-10,81%	-1,13%	11,53%	-14,30%	-1,38%	25,98%	-18,02%	3,98%	27,25%	-19,61%	3,82%	32,37%	-26,03%	3,17%
24/05/2021	7,68%	-9,82%	-1,07%	12,57%	-13,94%	-0,69%	28,15%	-18,84%	4,66%	30,33%	-20,39%	4,97%	32,21%	-24,05%	4,08%
25/05/2021	7,42%	-9,55%	-1,07%	13,37%	-13,16%	0,11%	28,42%	-17,56%	5,43%	31,38%	-19,05%	6,16%	31,57%	-24,50%	3,54%
26/05/2021	7,56%	-9,51%	-0,97%	14,82%	-12,56%	1,13%	29,68%	-16,18%	6,75%	33,57%	-17,34%	8,11%	32,79%	-22,95%	4,92%
27/05/2021	7,22%	-8,80%	-0,79%	16,16%	-12,02%	2,07%	30,99%	-15,59%	7,70%	35,85%	-16,76%	9,55%	31,29%	-21,94%	4,67%
28/05/2021	7,38%	-8,78%	-0,70%	17,10%	-11,87%	2,62%	32,63%	-15,70%	8,46%	37,59%	-16,87%	10,36%	31,41%	-21,52%	4,94%
29/05/2021	8,59%	-10,20%	-0,80%	12,32%	-13,13%	-0,40%	28,45%	-14,89%	6,78%	30,37%	-16,63%	6,87%	34,96%	-24,46%	5,25%
30/05/2021	8,07%	-9,61%	-0,77%	11,84%	-13,14%	-0,65%	26,06%	-15,28%	5,39%	27,97%	-17,26%	5,35%	33,26%	-24,57%	4,34%
31/05/2021	7,98%	-8,75%	-0,39%	13,67%	-11,53%	1,07%	26,48%	-14,48%	6,00%	29,77%	-16,03%	6,87%	32,67%	-21,88%	5,40%
01/06/2021	7,98%	-8,38%	-0,20%	14,81%	-11,59%	1,61%	30,11%	-14,41%	7,85%	33,69%	-15,91%	8,89%	33,05%	-21,42%	5,82%
02/06/2021	7,42%	-9,88%	-1,23%	16,64%	-12,24%	2,20%	32,28%	-13,81%	9,24%	37,35%	-14,80%	11,28%	32,66%	-20,92%	5,87%
03/06/2021	8,21%	-8,54%	-0,17%	17,45%	-10,34%	3,55%	31,65%	-13,32%	9,17%	36,61%	-13,93%	11,34%	34,66%	-19,98%	7,34%
04/06/2021	8,62%	-9,72%	-0,55%	16,55%	-12,42%	2,06%	32,10%	-16,23%	7,93%	35,91%	-17,30%	9,31%	34,68%	-23,12%	5,78%
05/06/2021	8,91%	-11,10%	-1,09%	13,82%	-13,67%	0,08%	28,47%	-16,23%	6,12%	30,73%	-17,53%	6,60%	35,92%	-24,74%	5,59%
06/06/2021	8,89%	-10,17%	-0,64%	14,70%	-12,31%	1,20%	28,08%	-14,63%	6,72%	31,24%	-15,67%	7,78%	35,09%	-22,85%	6,12%
07/06/2021	8,21%	-8,40%	-0,09%	17,89%	-10,27%	3,81%	32,89%	-12,65%	10,12%	38,35%	-13,31%	12,52%	33,34%	-20,02%	6,66%
08/06/2021	8,33%	-7,94%	0,20%	18,71%	-9,60%	4,56%	33,58%	-12,66%	10,46%	39,28%	-13,19%	13,05%	36,30%	-19,57%	8,36%
09/06/2021	8,55%	-8,53%	0,01%	19,18%	-10,56%	4,31%	33,93%	-14,58%	9,68%	39,50%	-15,21%	12,15%	38,43%	-19,83%	9,30%
10/06/2021	9,26%	-10,28%	-0,51%	18,36%	-12,47%	2,95%	32,89%	-17,56%	7,66%	37,61%	-18,22%	9,70%	36,91%	-22,56%	7,17%
11/06/2021	9,05%	-9,11%	-0,03%	19,94%	-10,94%	4,50%	33,20%	-14,71%	9,25%	39,52%	-15,20%	12,16%	35,41%	-20,82%	7,30%

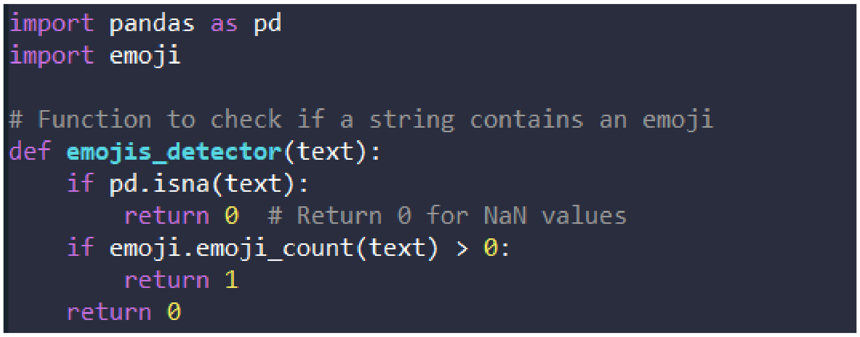
ATTACHED

Attached 1 - Comment extraction using PRAW



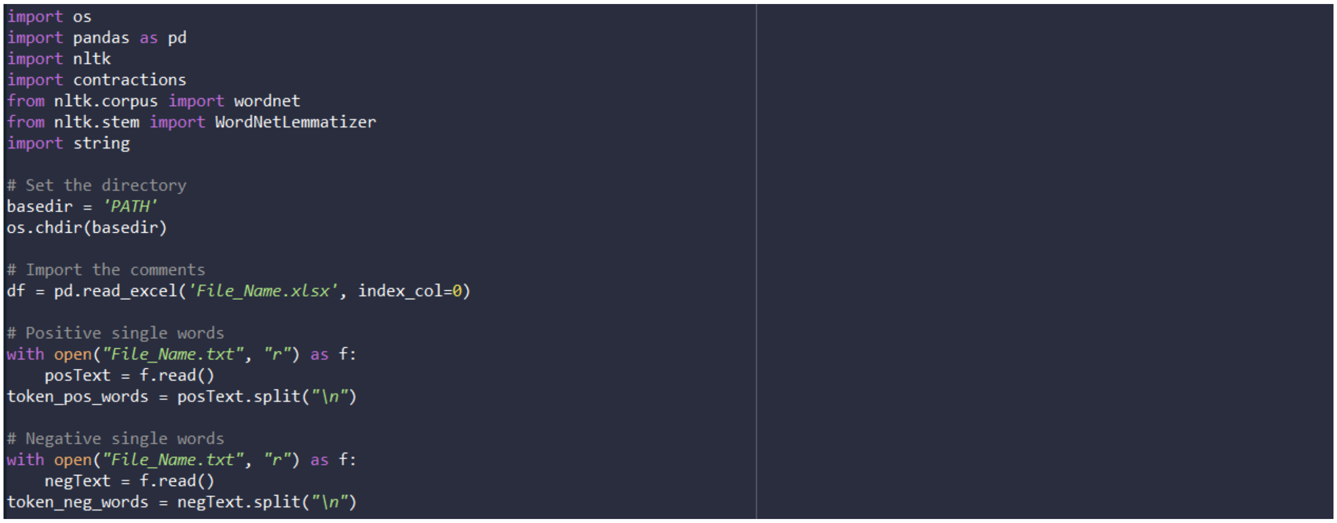
```
comm_list= []
url= 'INSERT URL'
submission = reddit.submission(url=url)

submission.comments.replace_more(limit=None)
for comment in submission.comments.list():
    comm_list.append([comment.created, comment.score, comment.body])
    df = pd.DataFrame(comm_list, columns=['Created', 'Score', 'Body'])
```

Attached 2 – Emojis detector

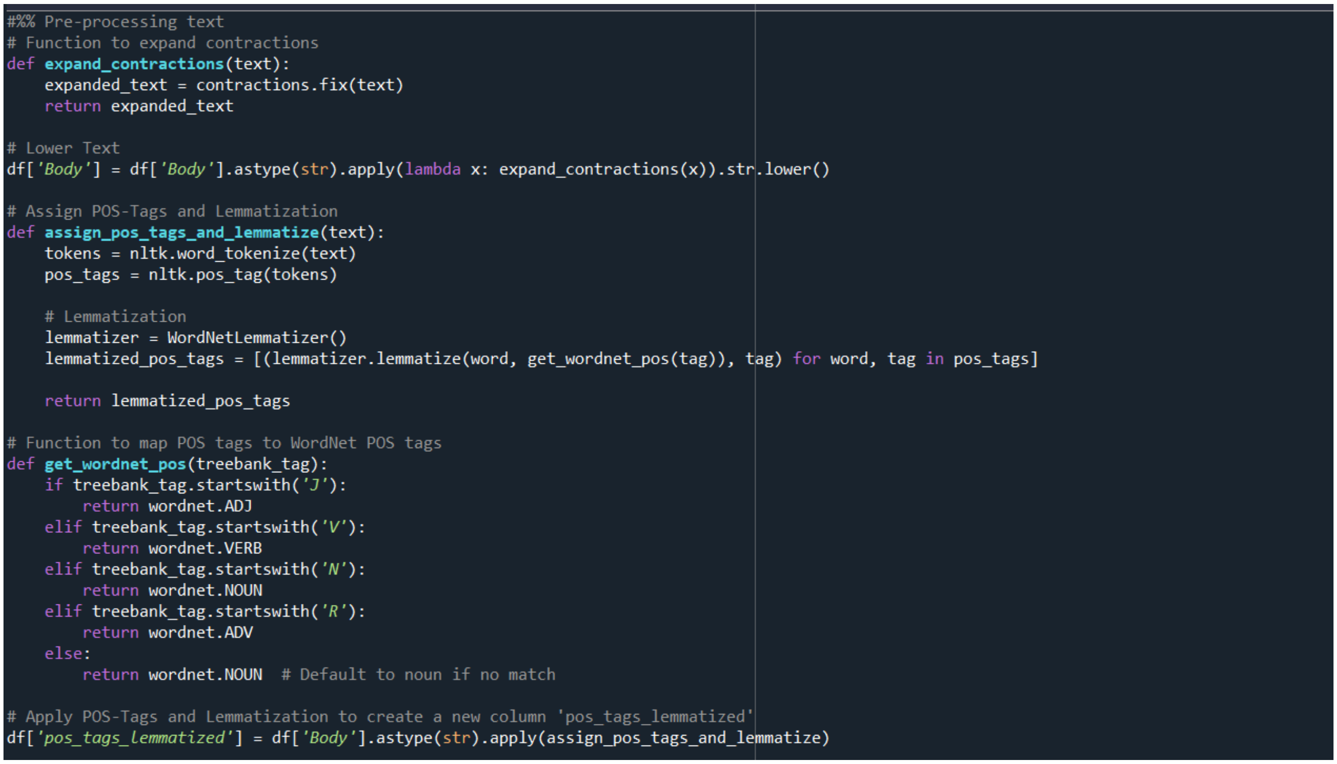


```
import pandas as pd
import emoji

# Function to check if a string contains an emoji
def emojis_detector(text):
    if pd.isna(text):
        return 0  # Return 0 for NaN values
    if emoji.emoji_count(text) > 0:
        return 1
    return 0
```

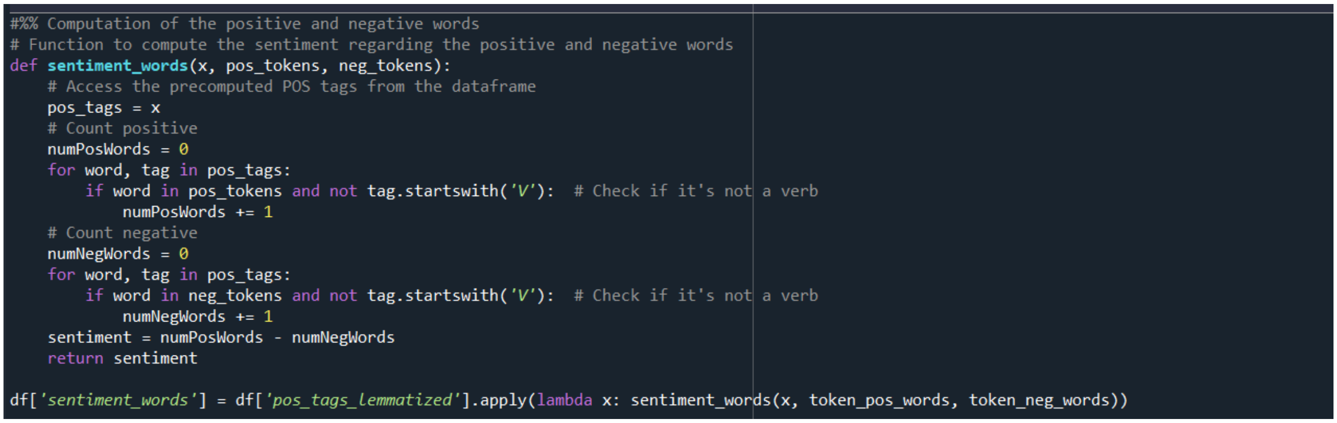
Attached 3 – Sentiment analysis



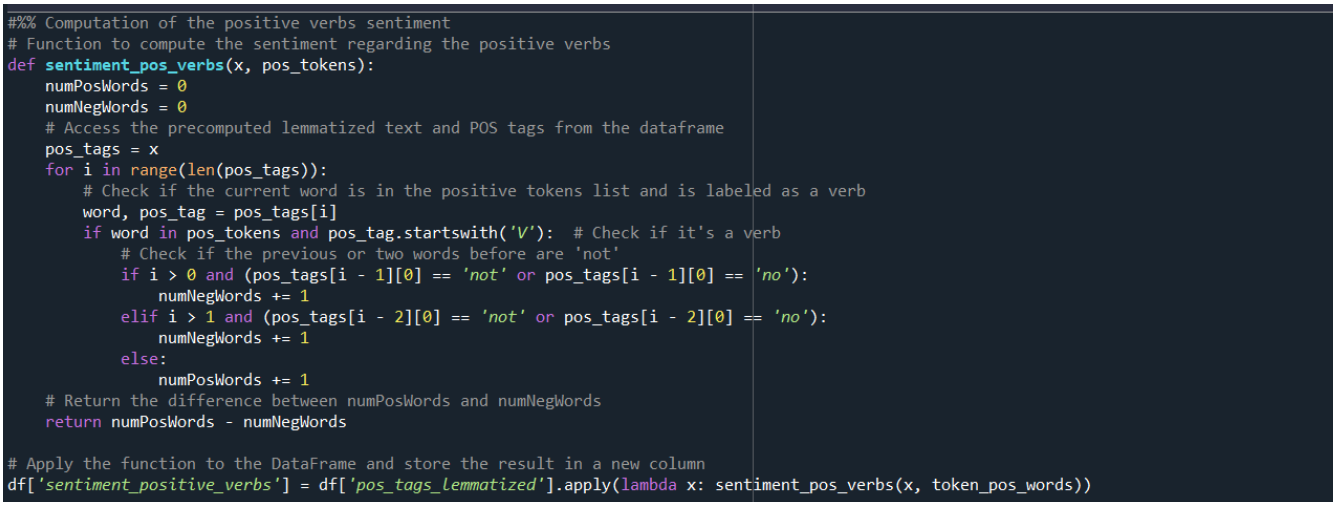
```
import os
import pandas as pd
import nltk
import contractions
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer
import string

# Set the directory
basedir = 'PATH'
os.chdir(basedir)

# Import the comments
df = pd.read_excel('File_Name.xlsx', index_col=0)

# Positive single words
with open("File_Name.txt", "r") as f:
    posText = f.read()
token_pos_words = posText.split("\n")

# Negative single words
with open("File_Name.txt", "r") as f:
    negText = f.read()
token_neg_words = negText.split("\n")
```



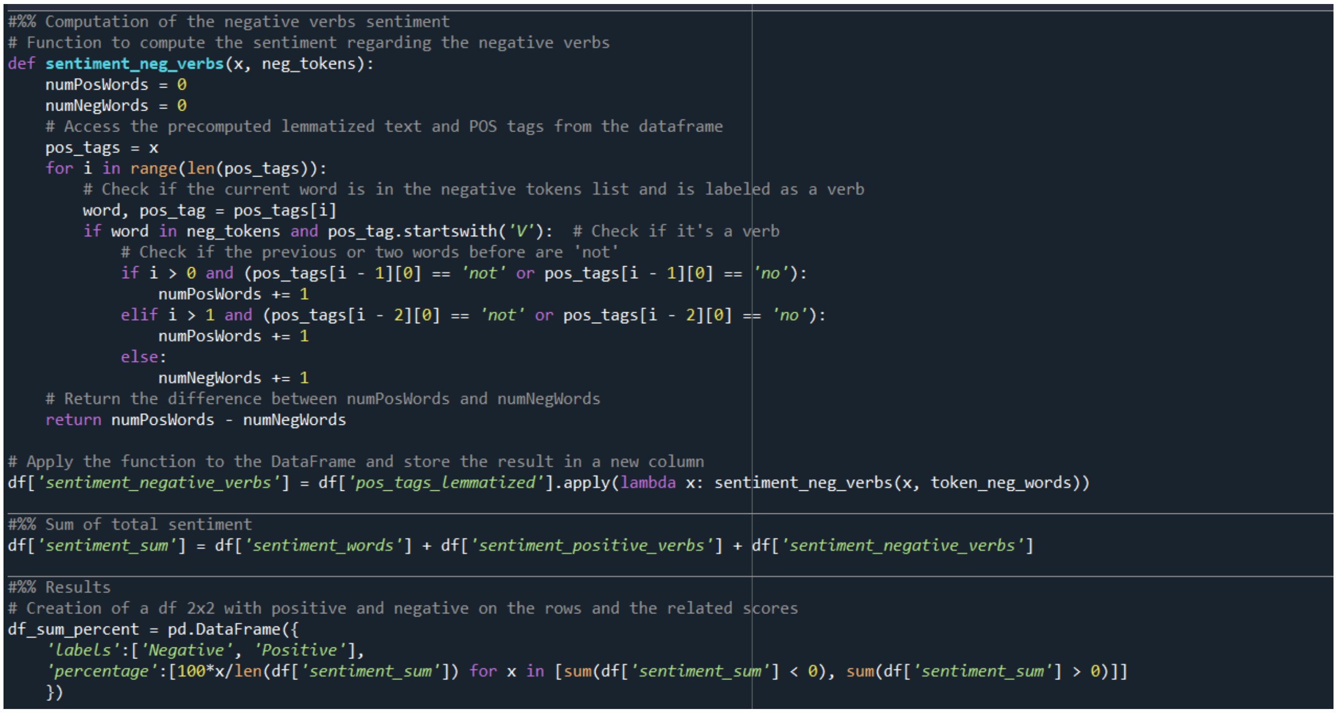
```python
#%% Pre-processing text
# Function to expand contractions
def expand_contractions(text):
    expanded_text = contractions.fix(text)
    return expanded_text

# Lower Text
df['Body'] = df['Body'].astype(str).apply(lambda x: expand_contractions(x)).str.lower()

# Assign POS-Tags and Lemmatization
def assign_pos_tags_and_lemmatize(text):
    tokens = nltk.word_tokenize(text)
    pos_tags = nltk.pos_tag(tokens)

    # Lemmatization
    lemmatizer = WordNetLemmatizer()
    lemmatized_pos_tags = [(lemmatizer.lemmatize(word, get_wordnet_pos(tag)), tag) for word, tag in pos_tags]

    return lemmatized_pos_tags

# Function to map POS tags to WordNet POS tags
def get_wordnet_pos(treebank_tag):
    if treebank_tag.startswith('J'):
        return wordnet.ADJ
    elif treebank_tag.startswith('V'):
        return wordnet.VERB
    elif treebank_tag.startswith('N'):
        return wordnet.NOUN
    elif treebank_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN  # Default to noun if no match

# Apply POS-Tags and Lemmatization to create a new column 'pos_tags_lemmatized'
df['pos_tags_lemmatized'] = df['Body'].astype(str).apply(assign_pos_tags_and_lemmatize)
```



```python
#%% Computation of the positive and negative words
# Function to compute the sentiment regarding the positive and negative words
def sentiment_words(x, pos_tokens, neg_tokens):
    # Access the precomputed POS tags from the dataframe
    pos_tags = x
    # Count positive
    numPosWords = 0
    for word, tag in pos_tags:
        if word in pos_tokens and not tag.startswith('V'):  # Check if it's not a verb
            numPosWords += 1
    # Count negative
    numNegWords = 0
    for word, tag in pos_tags:
        if word in neg_tokens and not tag.startswith('V'):  # Check if it's not a verb
            numNegWords += 1
    sentiment = numPosWords - numNegWords
    return sentiment

df['sentiment_words'] = df['pos_tags_lemmatized'].apply(lambda x: sentiment_words(x, token_pos_words, token_neg_words))
```



```python
#%% Computation of the positive verbs sentiment
# Function to compute the sentiment regarding the positive verbs
def sentiment_pos_verbs(x, pos_tokens):
    numPosWords = 0
    numNegWords = 0
    # Access the precomputed lemmatized text and POS tags from the dataframe
    pos_tags = x
    for i in range(len(pos_tags)):
        # Check if the current word is in the positive tokens list and is labeled as a verb
        word, pos_tag = pos_tags[i]
        if word in pos_tokens and pos_tag.startswith('V'):  # Check if it's a verb
            # Check if the previous or two words before are 'not'
            if i > 0 and (pos_tags[i - 1][0] == 'not' or pos_tags[i - 1][0] == 'no'):
                numNegWords += 1
            elif i > 1 and (pos_tags[i - 2][0] == 'not' or pos_tags[i - 2][0] == 'no'):
                numNegWords += 1
            else:
                numPosWords += 1
    # Return the difference between numPosWords and numNegWords
    return numPosWords - numNegWords

# Apply the function to the DataFrame and store the result in a new column
df['sentiment_positive_verbs'] = df['pos_tags_lemmatized'].apply(lambda x: sentiment_pos_verbs(x, token_pos_words))
```



```
#%% Computation of the negative verbs sentiment
# Function to compute the sentiment regarding the negative verbs
def sentiment_neg_verbs(x, neg_tokens):
    numPosWords = 0
    numNegWords = 0
    # Access the precomputed lemmatized text and POS tags from the dataframe
    pos_tags = x
    for i in range(len(pos_tags)):
        # Check if the current word is in the negative tokens list and is labeled as a verb
        word, pos_tag = pos_tags[i]
        if word in neg_tokens and pos_tag.startswith('V'):  # Check if it's a verb
            # Check if the previous or two words before are 'not'
            if i > 0 and (pos_tags[i - 1][0] == 'not' or pos_tags[i - 1][0] == 'no'):
                numPosWords += 1
            elif i > 1 and (pos_tags[i - 2][0] == 'not' or pos_tags[i - 2][0] == 'no'):
                numPosWords += 1
            else:
                numNegWords += 1
    # Return the difference between numPosWords and numNegWords
    return numPosWords - numNegWords

# Apply the function to the DataFrame and store the result in a new column
df['sentiment_negative_verbs'] = df['pos_tags_lemmatized'].apply(lambda x: sentiment_neg_verbs(x, token_neg_words))

#%% Sum of total sentiment
df['sentiment_sum'] = df['sentiment_words'] + df['sentiment_positive_verbs'] + df['sentiment_negative_verbs']

#%% Results
# Creation of a df 2x2 with positive and negative on the rows and the related scores
df_sum_percent = pd.DataFrame({
    'labels':['Negative', 'Positive'],
    'percentage':[100*x/len(df['sentiment_sum']) for x in [sum(df['sentiment_sum'] < 0), sum(df['sentiment_sum'] > 0)]]
    })
```

REFERENCES

BIBLIOGRAPHY

ASHBY, D., and JENSEN, C. T. (2018). *APIs for Dummies a Wiley brand.* John Wiley & Sons.

BARBIER, G., and LIU, H. (2011). *Data mining in social media*. Social Network Data Analytics. Springer, Boston, MA, pp. 327-352.

BENGFORT, B., BILBRO, R. and OJEDA, T. (2018). *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning.* O'Reilly Media.

BIRD, S., KLEIN, E., and LOPER, E. (2009). *Natural Language Processing with Python: Analysing Text with the Natural Language Toolkit.* O'Reilly Media.

HEISS, F., and BRUNNER, D. (2020). *Using Python for Introductory Econometrics*. Independently published. Düsseldorf, Germany.

SARKAR, D. (2016). *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data.* Independently published.

SHEPPARD, K. (2020). *Introduction to Python for Econometrics, Statistics and Data Analysis.* Independently published.

SCIENTIFICS ARTICLES

ALLEN, F., NOWAK, E., PIROVANO, M., and TENGULOV, A. (2021). *Squeezing shorts through social news platforms* (No. 21-31). Swiss Finance Institute.

ALOOSH, A., OUZAN, S., and SHAHZAD, S. J. H. (2022). *Bubbles across meme stocks and cryptocurrencies.* Finance Research Letters, 49.

ANAND, A., and PATHAK, J. (2021). *The role of Reddit in the GameStop short squeeze.* Economics Letters 211.

ASMI, A., and ISHAYA, T. (2012). *Negation identification and calculation in sentiment analysis.* The second international conference on advances in information mining and management (pp. 1-7).

BERNSTEIN, M., MONROY-HERNÁNDEZ, A., HARRY, D., ANDRÉ, P., PANOVICH, K., and VARGAS, G. (2011). *4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online*

Community. In Proceedings of the international AAAI conference on web and social media (Vol. 5, No. 1, pp. 50-57).

BOYLSTON, C., PALACIOS, B., TASSEV, P., and BRUCKMAN, A. (2021). *Wallstreetbets: positions or ban.* arXiv preprint arXiv:2101.12110.

CHEN, H., CHIANG, R. H., and STOREY, V. C. (2012). *Business intelligence and analytics: From big data to big impact.* MIS Quarterly, 36(4):1165-1188.

CHEN, H., DE, P., HU, Y. J., and HWANG, B. H. (2013). *Customers as advisors: The role of social media in financial markets.* Working paper.

CHEN, C. Y. H., DESPRÉS, R., GUO, L. and RENAULT, T. (2019). *What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble* (No. 2019-016). IRTG 1792 Discussion Paper.

CONRAD, F. G., GAGNON-BARTSCH, J. A., FERG, R. A., SCHOBER, M. F., PASEK, J., and HOU, E. (2021). *Social media as an alternative to surveys of opinions about the economy.* Social Science Computer Review, 39(4), 489-508.

CRUZ, R., KINYUA, J., and MUTIGWE, C. (2023). *Analysis of Social Media Impact on Stock Price Movements Using Machine Learning Anomaly Detection.* Intelligent Automation & Soft Computing, 36(3).

DAAS, P. J., PUTS, M. J., BUELENS, B., and HURK, P. A. V. D. (2015). *Big data as a source for official statistics.* Journal of Official Statistics, 31(2), 249-262.

DAVIS, A. K., PIGER, J. M., and SEDOR, L. M. (2012). *Beyond the numbers: Measuring the information content of earnings press release language.* Contemporary Accounting Research, 29(3), 845-868.

DICUONZO, G., GALEONE, G., ZAPPIMBULSO, E., and DELL'ATTI, V. (2019). *Risk management 4.0: The role of big data analytics in the bank sector.* International Journal of Economics and Financial Issues, 9(6), 40-47.

ENGELBERG, J. (2008). *Costly information processing: Evidence from earnings announcements.* In AFA 2009 San Francisco meetings paper.

GANDOMI, A. AND HAIDER, M. (2015). *Beyond the hype: big data concepts, methods, and analytics.* International Journal of Information Management, 35(2):137-144.

GENTZKOW, M., KELLY, B. T., and TADDY, M. (2019). *Text as Data.* Journal of Economic Literature, 57(3):535-74.

GUNDECHA, P., and LIU, H. (2012). *Mining Social Media: A Brief Introduction.* Tutorials in Operations Research, 1(4):1-17.

HEIDEMANN, J., KLIER, M., and PROBST, F. (2012). *Online social networks: A survey of a global phenomenon.* Computer Networks, 56(18):3866-3878.

HU, M., and LIU, B. (2004). *Mining and Summarizing Customer Reviews.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), Seattle, WA, 22–25 August.

HUTTO, C., and GILBERT, E. (2014). *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).

JEGADEESH, N., and WU, D. (2013). *Word power: A new approach for content analysis.* Journal of financial economics, 110(3), 712-729.

KEARNEY, C. and LIU, S. (2014). *Textual Sentiment in Finance: A Survey of Methods and Models.* International Review of Financial Analysis, 33:171–85.

LERMAN, K., and GHOSH, R. (2010). *Information contagion: An empirical study of the spread of news on digg and twitter social networks.* In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 4, No. 1, pp. 90-97).

LI, F. (2006). *Do stock market investors understand the risk sentiment of corporate annual reports?* SSRN Electronic Journal 898181.

LIU, B., and MCCONNELL, J. J. (2013). *The role of the media in corporate governance: Do the media influence managers' capital allocation decisions?* Journal of Financial Economics, 110(1), 1-17.

LOUGHRAN, T., and MCDONALD, B. (2011). *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-ks.* The Journal of Finance, 66(1):35–65.

LOUGHRAN, T., and MCDONALD, B. (2016). *Textual analysis in accounting and finance: A survey.* Journal of Accounting Research, 54(4):1187-1230.

LOUGHRAN, T., and MCDONALD, B. (2020). *Textual Analysis in Finance.* Annual Review of Financial Economics, 12(1):357-375.

LUCCHINI, L., AIELLO, L.M., ALESSANDRETTI, L., MORALES, G.D.F., STARNINI, M., and BARONCHELLI, A. (2021). *From Reddit to Wall Street: The role of committed minorities in financial collective action.* Royal Society Open Science 9(4), Article 211488.

LYÓCSA, S., BAUMÖHL, E. AND VÝROST, T. (2022). *YOLO trading: Riding with the herd during the GameStop episode.* Finance Research Letters, 46, Article 102359.

MANCINI, A., DESIDERIO, A., CLEMENTE, R. D., and CIMINI, G. (2021). *Self-induced emergence of consensus in social networks: Reddit and the GameStop short squeeze.* arXiv, 2112(07059).

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C., and HUNG BYERS, A. (2011). *Big data: The next frontier for innovation, competition, and productivity.* Report of the McKinsey Global Institute, McKinsey & Company.

NOBANEE, H. and ELLILI, N. O. D. (2023). *What do we know about meme stocks? A bibliometric and systematic review, current streams, developments, and directions for future research.* International Review of Economics & Finance, 85(C):589-602.

NONNECKE, B., and PREECE, J. (2000). *Lurker demographics: Counting the silent.* In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 73-80).

NOVAK, P. K., SMAILOVIĆ, J., SLUBAN, B. AND MOZETIČ, I. (2015). *'Sentiment of emojis'*, Plos One 10:(12): e0144296.

PAGOLU, V. S., REDDY, K. N., PANDA, G., and MAJHI, B. (2016). *Sentiment analysis of Twitter data for predicting stock market movements.* In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.

PANG, B., and LEE, L. (2008). *Opinion mining and sentiment analysis.* Foundations and Trends® in information retrieval, 2(1-2), 1-5.

POLANYI, L., and ZAENEN, A. (2006). *Contextual valence shifters.* Computing attitude and affect in text: Theory and applications, 1-10.

PREIS, T., MOAT, H. S., and STANLEY, H. E. (2013). *Quantifying Trading Behaviour in Financial Markets Using Google Trends.* Scientific Reports, 3(1):1-6.

RENAULT, T. (2017). *Intraday online investor sentiment and return patterns in the U.S. stock market.* Journal of Banking & Finance, 84(C):25-40.

SAGGI, M. K., and JAIN, S. (2018). *A survey towards an integration of big data analytics to big insights for value-creation.* Information Processing & Management, 54(5), 758-790.

SAGIROGLU, S., and SINANC, D. (2013). *Big data: A review.* In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE.

SMITH, B. K., and GUSTAFSON, A. (2017). *Using Wikipedia to predict election outcomes: Online behavior as a predictor of voting.* Public Opinion Quarterly, 81, 714-735.

TAUSCZIK, Y. R., & PENNEBAKER, J. W. (2010). *The psychological meaning of words: LIWC and computerized text analysis methods.* Journal of language and social psychology, 29(1), 24-54.

TETLOCK P. C. (2007). *Giving content to investor sentiment: the role of media in the stock market.* The Journal of finance, 62(3):1139-1168.

TETLOCK, P. C., SAAR-TSECHANSKY, M., and MACSKASSY, S. (2008). *More than words: Quantifying language to measure firms' fundamentals.* The journal of finance, 63(3):1437-1467.

WESTERMAN, D., SPENCE, P. R., and VAN DER HEIDE B. (2014). *Social media as information source: Recency of updates and credibility of information.* Journal of Computer-Mediated Communication, 19 (2):171-183.

ONLINE SOURCES

AGENDA DIGITALE, 2018. *Il Contesto nei sistemi informativi: cos'è e perché è sempre più importante* [online]. Agenda Digitale: <https://www.agendadigitale.eu/cultura-digitale/il-contesto-nei-sistemi-informativi-cose-e-perche-e-sempre-piu-importante/> [05/10/2023]

ARCHIVE, 2014. *About the Internet Archive* [online]. Internet Archive: <https://archive.org/about/> [01/02/2024]

ARCHIVE, 2022a. *Archive.org page overview* [online]. Internet Archive: <https://help.archive.org/help/archive-org-page-overview/> [13/06/2023]

ARCHIVE, 2022b. *Save pages in-the Wayback Machine* [online]. Internet Archive: <https://help.archive.org/help/save-pages-in-the-wayback-machine/> [13/06/2023]

ARCHIVE, 2023. *Search – A Basic Guide* [online]. Internet Archive: <https://help.archive.org/help/search-a-basic-guide/> [13/06/2023]

ASHENFELDER, M. 2011. *The Average Lifespan of a Webpage* [online]. Library of Congress Blogs: <https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-webpage/> [01/02/2024]

BUSINESS ENGLISH, 2024. *r/wallstreetbets vocabulary* [online]. Businessenglish: <https://www.businessenglish.com.hk/blog/financial-english-blog/r-wallstreetbets-vocabulary.html> [14/01/2024]

CLARK, M., 2021. *New research shows how many important links on the web get lost to time* [online]. <https://www.theverge.com/2021/5/21/22447690/link-rot-research-new-york-times-domain-hijacking> [21/11/2023]

COMPUTER SCIENCE, 2017. *High and Low Level Languages* [online]. Computerscience: <https://www.computerscience.gcse.guru/theory/high-low-level-languages> [14/12/2023]

CONGRUITY 360, 2023. *The Future of Data: Unstructured Data Statistics You Should Know* [online]. Congruity 360: <https://www.congruity360.com/blog/the-future-of-data-unstructured-data-statistics-you-should-know/> [01/02/2024]

COUGHLIN, T., 2018. *175 Zettabytes By 2025* [online]. Forbes:

<https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/> [01/02/2024]

CUKIER, K. (2010). *Data, data everywhere: A special report on managing information.* The Economist. [online]. SMU: <https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://s2.smu.edu/tfomby/eco5385_eco6380/The%2520Economist-data-data-everywhere.pdf&ved=2ahUKEwiek-aU6KWFAxXx_7sIHZdXDn0QFnoECBgQAQ&usg=AOvVaw19NUD5Ri0-BsB2D3j95u6l > [01/02/2024]

DAILEY, N., 2021. *Wall Street Bets claimed the top 3 most popular posts on all of Reddit this year. Here's what they said* [online]. Business Insider: <https://markets.businessinsider.com/news/stocks/wall-street-bets-top-reddit-posts-in-2021-2021-12> [01/02/2024]

DEAN, B., 2023. *Reddit User and Growth Stats (Updated March 2023)* [online]. Backlinko: <https://backlinko.com/reddit-users> [30/01/2024]

DEVINTERFACE, 2023. *Everything you need to know about APIs in software architecture* [online]. Devinterface: <https://www.devinterface.com/en/blog/everything-you-need-to-know-about-apis-in-software-architecture> [23/11/2023]

DIALANI, P., 2020. *The Future of Data Revolution will be Unstructured Data* [online]. Analytics Insight: <https://www.analyticsinsight.net/the-future-of-data-revolution-will-be-unstructured-data/> [30/01/2024]

DIXON, M.J., 2022a. *Frequency of Reddit use in the United States as of 3rd quarter 2020* [online]. Statista: <https://www.statista.com/statistics/815177/reddit-usage-frequency-usa/> [30/01/2024]

DIXON, M.J., 2022b. *Percentage of internet users who use Reddit in the United Kingdom (UK) as of 3rd quarter 2020, by age group* [online]. Statista: <https://www.statista.com/statistics/1184024/reddit-user-demographics/> [30/01/2024]

DIXON, M.J., 2022c. *Percentage of U.S. adults who use Reddit as of February 2021, by age group* [online]. Statista: <https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/> [30/01/2024]

DIXON, M.J., 2022d. *Percentage of U.S. adults who use Reddit as of February 2021, by annual household income* [online]. Statista: <https://www.statista.com/statistics/261774/share-of-us-internet-users-who-use-reddit-by-annual-income/> [30/01/2024]

DIXON, M.J., 2022e. *Percentage of U.S. adults who use Reddit as of February 2021, by education level* [online]. Statista: <https://www.statista.com/statistics/261776/share-of-us-internet-users-who-use-reddit-by-education-level/> [30/01/2024]

DIXON, M.J., 2022f. *Percentage of U.S. adults who use Reddit as of February 2021, by urbanity* [online]. Statista: <https://www.statista.com/statistics/261783/share-of-us-internet-users-who-use-reddit-by-urbanity/> [30/01/2024]

DIXON, M.J., 2023. *Average daily active users (DAU) of Reddit in June 2021 and December 2022* [online]. Statista: <https://www.statista.com/statistics/1324264/reddit-daily-active-users/> [30/01/2024]

DIXON, M.J., 2023b. *Distribution of Reddit users worldwide as of 3rd quarter 2022, by gender* [online]. Statista: <https://www.statista.com/statistics/1255182/distribution-of-users-on-reddit-worldwide-gender/> [30/01/2024]

DUPREY, R., 2021. *Why FuboTV Stock Was 17% Higher Today* [online]. The Motley Fool: <https://www.fool.com/investing/2021/06/02/why-fubotv-is-soaring-15-higher-today/> [10/03/2023]

ELIAHU, R., 2023. *The Power of reddit's PRAW* [online]. Medium: <https://medium.com/@Eliahu.ran/the-power-of-reddits-praw-34c020526388> [14/06/2023]

EMOJI COMBOS, 2024. *Wall Street Bets Emojis & Text* [online]. Emoji Combos: <https://emojicombos.com/wall-street-bets> [15/01/2024]

EMOJI PARTY, 2024. *wallstreetbets emoji collection* [online]. Emoji party: <https://emoji.party/wallstreetbets> [14/01/2024]

EMOJIPEDIA, 2024. *Emojipedia* [online]. Emojipedia: <https://emojipedia.org/> [14/01/2024]

EPU, 2012. *US EPU (Monthly, Daily, Categorical)* [online]. Economic Policy Uncertainty: <https://www.policyuncertainty.com/us_monthly.html> [02/01/2024]

EUROSTAT, 2024. *Experimental statistics. Overview* [online]. Eurostat: <https://ec.europa.eu/eurostat/web/experimental-statistics/overview> [30/01/2024]

EVEREST GROUP, 2021. *Intelligent Document Processing (IDP) Adoption Swells as Enterprises Seek to Lower Costs Through Automation; IDP Market to Grow 55-65% in Next Year | Press Release* [online]. Everest Group: <https://www.everestgrp.com/2021-07-intelligent-document-processing-idp-adoption-swells-as-enterprises-seek-to-lower-costs-through-automation-idp-market-to-grow-55-65-in-next-year-press-release-.html> [01/02/2024]

FASTWEB, 2022. *Cos'è Reddit e come funziona* [online]. Fastweb: <https://www.fastweb.it/fastweb-plus/digital-marketing-social/cose-reddit-e-come-funziona/> [30/01/2024]

GOT API, 2023. *What Is an API Wrapper* [online]. GotApi: <https://gotapi.com/what-is-an-api-wrapper/> [14/11/2023]

GRAHAM, M., 2016. *FAQs for some new features available in the Beta Wayback Machine* [online]. Internet Archive Blogs: <https://blog.archive.org/2016/10/24/faqs-for-some-new-features-available-in-the-beta-wayback-machine/> [14/06/2023]

GRAHAM, M., 2017. *Wayback Machine Playback… now with Timestamps!* [online]. Internet Archive Blogs: <https://blog.archive.org/2017/10/05/wayback-machine-playback-now-with-timestamps/> [15/06/2023]

HOOTSUITE, 2020. *Thread* [online]. Hootsuite: <https://blog.hootsuite.com/social-media-definitions/thread/> [03/11/2023]

IBM, 2023. *What is an API?* [online]. IBM: <https://www.ibm.com/topics/api> [22/08/2023]

IONOS, 2019. *Tra notizie e meme: Reddit* [online]. Digital Guide IONOS: <https://www.ionos.it/digitalguide/online-marketing/social-media/cose-reddit-e-come-funziona/> [30/01/2024]

ISAAC, M., 2023. *Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems* [online]. The New York Times: <https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html> [15/06/2023]

ISTAT, 2023a. *Experimental statistics* [online]. Istat: <https://www.istat.it/en/experimental-statistics> [30/01/2024]

ISTAT, 2023b. *Social Mood on Economy Index* [online]. Istat: <https://www.istat.it/en/archivio/219600> [30/01/2024]

KULKARNI, R., 2019. *Big Data Goes Big* [online]. Forbes: <https://www.forbes.com/sites/rkulkarni/2019/02/07/big-data-goes-big/?sh=2fc64be920d7> [01/02/2024]

LA MONICA, P., 2021. *Wendy's stock surges as Reddit crowd talks up 'chicken tendies'* [online]. CNN Business: <https://edition.cnn.com/2021/06/08/investing/wendys-reddit-wallstreeetbets/index.html> [10/03/2023]

LIN, Y., 2023. *10 Reddit Statistics Every Marketer Should Know in 2024 [Infographic]* [online]. Oberlo: <https://www.oberlo.com/blog/reddit-statistics> [30/01/2024]

LOPS, V., 2023. *L'incredibile statistica di Wall Street: da 95 anni scende il lunedì e guadagna martedì e mercoledì* [online]. Il Sole 24 Ore: <https://www.ilsole24ore.com/art/l-incredibile-statistica-wall-street-95-anni-scende-lunedi-e-guadagna-martedi-e-mercoledi-AEFBZsSD> [10/01/2024]

LOUGHRAN, T., and MCDONALD, B., 2022. *Loughran-McDonald Master Dictionary w/ Sentiment Word Lists* [online]. University of Notre Dame: <https://sraf.nd.edu/loughranmcdonald-master-dictionary/> [15/11/2023]

LUTKEVICH, B., 2022. *Application programming interface (API)* [online]. TechTarget: <https://www.techtarget.com/searchapparchitecture/definition/application-program-interface-API> [25/07/2023]

LYDEN, C., 2017. *How Do Search Engines Work?* [online]. Call Raid: <https://www.callrail.com/blog/search-engines-work> [25/09/2023]

MANAW, P., 2023. *Wrapper a concept, Use cases, and Drawbacks* [online]. Medium: <https://medium.com/@pmmanav/why-and-where-to-use-wrapper-f49559e26d73> [14/06/2023]

METRICS FOR REDDIT, 2023. *New subreddits by month (How Reddit grew over time)* [online]. <https://frontpagemetrics.com/month/> [30/01/2024]

MIAO, H., STEVENS, P., 2021. *Clean Energy Fuels stock soars more than 30% as retail traders pick new targets* [online]. CNBC: <https://www.cnbc.com/2021/06/09/clean-energy-fuels-contextlogic-soar-as-retail-traders-pick-new-meme-stocks.html> [10/03/2023]

MORRIS, W., 2023. *The Basics of How Search Engine Indexing Works* [online]. Elegant themes: <https://www.elegantthemes.com/blog/wordpress/how-search-engine-indexing-works> [25/09/2023]

OHANIAN, A., 2013. *Reddit co-founder: How to turn failure into fuel* [online]. USA Today Opinion: <https://eu.usatoday.com/story/opinion/2013/10/17/millennials-ohanion-reddit-column/3004033/> [01/02/2024]

PELLICCIA, F., 2020. *Come sapere quando è stata pubblicata una pagina web* [online]. Francesco Pelliccia: <https://www.francescopelliccia.it/come-sapere-quando-e-stata-pubblicata-una-pagina-web/> [10/03/2023]

PRAW, 2023a. *PRAW: The Python Reddit API Wrapper* [online]. PRAW: <https://praw.readthedocs.io/en/stable/> [08/03/2023]

PRAW, 2023b. *Comment Extraction and Parsing* [online]. PRAW: <https://praw.readthedocs.io/en/stable/tutorials/comments.html> [08/03/2023]

RAIS, R., 2022. *Link rotti: cosa sono e quali sono gli effetti* [online]. OUTOFBIT: <https://www.outofbit.it/link-rotti-approfondimenti/> [21/11/2023]

RAPIDAPI, 2023. *What is an API? API definition* [online]. Rapid API: <https://rapidapi.com/blog/api-glossary/api/> [22/08/2023]
<https://web.archive.org/web/20230724170631/https://blog-proxy.rapidapi.com/api-glossary/api/>

REDDIT, 2017. *r/wallstreetbets terminology guide* [online]. Reddit: <https://www.reddit.com/r/wallstreetbets/comments/7948ov/rwallstreetbets_terminology_guide/> [25/01/2024]

REDDIT, 2021. *Basic guide to Wallstreetbets culture for Newcomers* [online]. Reddit: <https://www.reddit.com/r/wallstreetbets/comments/l7fr21/basic_guide_to_wallstreetbets_culture_for/> [16/01/2024]

REDDIT, 2023. *Api - reddit.com* [online]. <https://www.reddit.com/wiki/api/> [08/03/2023]

REDDIT, 2024. *Reddit* [online]. Reddit: <https://www.redditinc.com/> [30/01/2024]

RENAULT, T., 2022. *Data & Database* [online]. <https://www.thomas-renault.com/data.php> [30/11/2023]

ROUSE, 2015a. *Broken Link* [online]. Techopedia: <https://www.techopedia.com/definition/23236/broken-link> [21/11/2023]

ROUSE, M., 2015b. *Indexing* [online]. Techopedia: <https://www.techopedia.com/definition/7705/indexing> [18/10/2023]

ROUSE, M., 2016. *Fortran* [online]. Techopedia: <https://www.techopedia.com/definition/24111/fortran> [14/12/2023]

ROUSE, M., 2017. *Web Crawler* [online]. Techopedia: <https://www.techopedia.com/definition/10008/web-crawler> [18/10/2023]

ROUSE, M., 2022. *Python* [online]. Techopedia: <https://www.techopedia.com/definition/3533/python> [14/12/2023]

SCOTT, G., 2022. *Application Programming Interface (API): Definition and Examples* [online]. Investopedia: <https://www.investopedia.com/terms/a/application-programming-interface.asp> [25/07/2023]

SCOWL, 2016. *Version 6 of the 12dicts word lists* [online]. SCOWL (And Friends): <http://wordlist.aspell.net/12dicts-readme/> [18/12/2023]

SIMILARWEB, 2023. *reddit.com Traffic & Engagement Analysis* [online]. Similarweb: <https://www.similarweb.com/website/reddit.com/#traffic> [30/01/2024]

STOCKTWITS, 2023. [online]. <https://stocktwits.com/> [21/12/2023]

SYMBL, 2024. *Miscellaneous Symbols and Pictographs* [online]. Unicode Character Table: <https://symbl.cc/en/unicode/blocks/miscellaneous-symbols-and-pictographs/> [25/01/2024]

TECHTERMS, 2006. *Spider* [online]. TechTerms: <https://techterms.com/definition/spider> [18/10/2023]

TECHTERMS, 2008. *Index* [online]. TechTerms: <https://techterms.com/definition/index> [18/10/2023]

TECHTERMS, 2016a. *API* [online]. TechTerms: <https://techterms.com/definition/api> [22/08/2023]

TECHTERMS, 2016b. *Python* [online]. TechTerms: <https://techterms.com/definition/python> [14/12/2023]

TECHTERMS, 2019. *Wrapper* [online]. TechTerms: <https://techterms.com/definition/wrapper> [12/09/2023]

TECHTERMS, 2021. *Timestamp* [online]. TechTerms: <https://techterms.com/definition/timestamp> [18/10/2023]

TITINNANZI, E., 2018. *Come trovare il giorno in cui è stata pubblicata una pagina web* [online]. Idee per computer ed internet: <https://www.ideepercomputeredinternet.com/2018/04/cercare-data-pubblicazione-post.html> [10/03/2023]

WEISS, R., 2003. *On the Web, Research Work Proves Ephemeral* [online]. The Washington Post:

<https://washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/> [01/02/2024]

WHITE, W., 2021. *CLOV Stock: 9 Things to Know About Meme Stock Favorite Clover Health as Shares Surge* [online]. Nasdaq: <https://www.nasdaq.com/articles/clov-stock:-9-things-to-know-about-meme-stock-favorite-clover-health-as-shares-surge-2021> [10/03/2023]

WOOLARD, C., 2023. *Making sense of the rise of unstructured data* [online]. Automationhero: <https://automationhero.ai/blog/making-sense-of-the-rise-of-unstructured-data/> [01/02/2024]