Data Extraction and Sentiment Analysis of Social Media

Kshitij Sekhar Dutta
Student
Department of Computer Science and Engineering
Manipal University Jaipur Jaipur, India

ABSTRACT
Social media nowadays has become synonymous with the internet. Social media platforms have long evolved from being simple forums where people could post photos and thoughts to now being a base where people can launch successful entrepreneurial businesses or even turn into influencers, potentially earning them millions. This paper aims to harness the power of Natural Language Processing through Sentiment Analysis and apply it to one of the most popular social media forums right now, Reddit. Reddit has 73.1 million daily active users and 267.5 million weekly active users. There are more than 100,000 active subreddits (sub-forums) on the platform.

This paper utilizes Reddit APIs to employ a crawler that scrapes data from Reddit and orders them into a single data set. Then, the paper examines the structure of this data set. Through this data set it then analyses what the current topics of discussions were about, what the perceived opinions of the users were about the various topics. This paper intends to find if there is a correlation between the amount and type of emotions. This paper also highlights the concerns with using Sentiment Analysis and some other applications of it in real-life.

General Terms
Natural Language Processing

Keywords
Sentiment Analysis, Data Mining, Natural Language Processing, Social Media, Reddit

1. INTRODUCTION
Reddit is a popular user driven forum that consists of many communities called subreddits dedicated to discussing a pre-defined topic. Users can submit original content or links to other sites and have discussions. This is a unique social network because the emphasis is on the community rather than a single user. Analysed data regarding discussions about political trends, events and other policy decisions. Scraping the data using APIs provided by Reddit and applying data mining and sentiment analysis techniques on the scraped data was done. Through this analysis, the paper provides a clear insight into how users of such sites behave and interact with one another while also gaining insights into their opinions and biases towards various issues. Reddit is a site that receives a tremendous amount of traffic, millions of users a day. To further explain how Reddit works, each post has an associated comment thread, and users of Reddit can vote the comments up (upvote) or down (downvote), generating a net score, or "Karma," for each comment. The Karma gained from upvoting or downvoting is not directly proportional to it. Users aspire to collect this "Karma," and these comments build the community on Reddit. When reading the content of the comment thread, however, it is often unclear why some comments succeed and receive high Karma while other comments lose Karma. This project wishes to get better insight into Reddit communities through the sentiment analysis of Reddit comments and on Reddit comments to characterize the voting patterns of Reddit users and determines the Karma strength of Reddit comments through identifying comments with positive and negative Karma.

The ideal opinion-mining tool would be to process a set of search results for a given item, generating a list of product attributes (quality features, etc.) and aggregating opinions about each of them (poor, mixed, good). However, the term has recently also been interpreted more broadly to include many different types of analysis of the evaluative text. In general, opinions can be expressed on anything, e.g., a product, a service, a topic, an individual, an organization, or an event. The general term object is used to denote the entity that has been commented on. Thus, object O can be defined as an entity which can be a product, topic, person, event, or organization. It is associated with a pair, O: (T, A), where T is a hierarchy or taxonomy of components (or parts) and sub-components of O, and A is a set of attributes of O. Each component has its own set of sub-components and attributes. The word features are used to represent both components and attributes. For an evaluative document D, opinion passage on a feature f of the object O evaluated in D is a group of consecutive sentences in D that expresses a positive or negative opinion on f. [1] Research on opinion mining or sentiment analysis started with identifying opinion (or sentiment) bearing words, e.g., great, amazing, wonderful, bad, and poor. Many researchers have worked on mining such words and identifying their semantic orientations or polarity determination (i.e., positive, negative and neutral). The researchers have identified several linguistic rules that can be exploited to identify opinion words and their orientations from a large corpus. In a theory, a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet.

The history of the phrase sentiment analysis parallels that of ―opinion mining in certain respects. The term―sentimental used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments that appear in 2001 paper by Das and Chen. Subsequently, this concept was adopted and enhanced by Turney and Pang in 2005 [2]. These events together may explain the popularity of―sentiment analysis among communities self-identified as focused on NLP.

In particular, sentiment analysis on online reviews has become a hot research field. Studies on sentiment analysis mainly focus on framework and lexicon construction, feature extraction, and polarity determination. This review conducts an overall survey of the three major research fields in sentiment analysis: framework, feature extraction and sentiment analysis, making a summary and analysis of the present development, and giving a detailed introduction of its application in business and Blogs. Despite the current immaturity of related research, sentiment analysis of online review has taken its position as an emerging

research frontline, which takes advantage of the achievements in many areas, such as text mining, natural language processing, web mining, and machine learning. But the related research did not take place until recently, and semantic parsing and understanding exhibit high complexity, the overall research in this field being in its infantry. Still a lot of problems need further exploration and solution as follows:

(1) Insufficient empirical language data and platform.

(2) No breakthrough in textual sentiment analysis.

(3) No research on the commercial value of online product reviews.

While there has been a fair amount of research on how sentiments are expressed in genres such as online reviews and news articles, how sentiments are expressed given the informal language and message-length constraints of microblogging has been much less studied. Features such as automatic part-of-speech tags and resources such as sentiment lexicons have proved useful for sentiment analysis in other domains, but will they also prove useful for sentiment analysis in Reddit? This paper begins to investigate this question. Another challenge of microblogging is the incredible breadth of topic that is covered. It is not an exaggeration to say that people post about anything and everything.

The growing expansion of contents, placed on the Web, provides a huge collection of textual resources. People share their experiences, opinions or simply talk just about whatever concerns them online. A large amount of available data attracts system developers, studying opinion mining and analysis. In this paper, the primary and underlying idea is that the fact of knowing how people feel about certain topics can be considered as a classification task. People's feelings can be positive, negative or neutral. A sentiment is often represented in subtle or complex ways in a text. An online user can use a diverse range of other techniques to express his or her emotions. Apart from that, s/he may mix objective and subjective information about a certain topic. On top of that, data gathered from the World Wide Web often contain a lot of noise. Indeed, the task of automatic sentiment recognition in the online text becomes more difficult for all the aforementioned reasons.

2. LITERATURE REVIEW

A There have been many papers written on sentiment analysis for the domain of blogs and product reviews. Pang [2] gives a survey of sentiment analysis. Researchers have also analyzed the brand impact of microblogging. Overall, text classification using machine learning is a well-studied field. Pang also researched the effects of various machine learning techniques such as Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) in the specific domain of movie reviews. They were able to achieve an accuracy of 82.9% using SVM and a unigram model. Researchers have also worked on detecting sentiment in text. Choi [3] presents a simple algorithm, called semantic orientation, for detecting sentiment. Pang also presents a hierarchical scheme in which text is first classified as containing sentiment, and then classified as positive or negative.

Work has been done in using emoticons as labels for positive and sentiment. This is very relevant to Reddit because many users have emoticons in their posts and comments. Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level. There are many established methods for sentiment analysis at the sentence and paragraph level. In T. Mullen's work [4] the authors discussed the application of support vector machines in sentiment analysis with diverse information source. In Pang's work, the authors applied minimum cuts in graphs to extract the subjective portion of texts they were studying and used machine learning methods to perform sentiment analysis on those snippets of texts only. In T. Wilson's work [5] the authors discussed categorizing texts into polar and neutral first before determining whether a positive or negative sentiment is expressed through the text. However, in N. Godbole's work [6], the authors operate on the premise that little neutrality exists in online texts. In the work [6], the authors developed techniques that algorithmically identify large number (hundreds) of adjectives, each with an assigned score of polarity, from around a dozen seed adjectives. Their methods expand two clusters of adjectives (positive and negative word groups) by recursively querying the synonyms and antonyms from WordNet. Since recursive search quickly connects words from the two clusters, they implemented several precautionary Data Extraction and Mining of Reddit User Data 5 measures such as assigning weights which decrease exponentially as the number of hops increases. This confirms that the algorithm-generated adjectives are highly accurate by comparing them to the results of manually picked word lists. It is worth pointing out that this work uses Lydia as the backbone to process a large amount of news and blogs. [7]

3. METHODOLOGY

3.1 Creating a Crawler

The first step in this project is to build a simple crawler that scrapes data from Reddit. This will be done using PRAW in Python. For a quick definition of these frameworks, Python is a high-level programming language for general-purpose program writing, created by Guido van Rossum and was released in 1991. An interpreted language, Python has a design idea that stresses code readability (especially using whitespace indentation to bound code blocks rather than curly brackets or keywords), and a grammar that allows programmers to express concepts in less lines of code than might be used in languages such as Java or C++. Meanwhile, PRAW, a shortening for 'Python Reddit API Wrapper', is a python package that enables for simple access to Reddit's API. PRAW, a shortening for "Python Reddit API Wrapper", is a python package that permits simple access to Reddit's API. PRAW targets to be easy to use and follows all of Reddit's API rules. PRAW has enabled the effective data mining and scraping of Reddit in a licensed way so that the bots do not adversely affect the site and the developers can work on their projects for analysis and sentiments extraction.

To scrape Reddit, we will need to obtain a valid API key from them. The key was first generated on the Reddit profile which gives the access to the crawler bot to crawl the Reddit page legally while binding to all the rules and regulations of the site in order that the sites working is not affected. It will be done as follows:

Step 1: Create a reddit account.

Step 2: Navigate to preferences → apps.

Step 3: Get the client id and client secret keys by filling in the details.

Step 4: Create an init file with the following contents: Reddit username, password, client id, client

secret

Step 5: Create an empty file called pandc.txt

Step 6: Authenticate:

Step 6.1: Read data from init file.

Step 6.2: Build json object and pass data to reddit.

Step 6.3: Authenticate the bot and store the returned data in an object.

Step 7: Fetch data:

Step 7.1: Get "hot" (popular right now) posts and comments from the subreddit 'India'. (This subreddit will be the main focal point of this research).

Step 7.2: If post or comment text matches a regex of an interested topic,

Step 7.2.1: Save post or comment id.

Step 7.2.2: Add id to pandc text file via file handling.

Step 7.2.3: Save body.

Step 7.2.4: Else, go to the next comment.

Step 8: Set schedule:

Step 8.1: To work around Reddit API limits, set scraping schedule to few minutes.

Step 8.2: Call fetch data

Step 8.3: Wait for a few seconds.

Step 8.4: Repeat

Step 9: Run analysis:

Step 9.1: Get total comments parsed.

Step 9.2: Get total comments that are relevant.

Step 9.3: Get the percentage of relevant comments.

Step 9.4: Run analysis on the data collected.

3.2 Running analysis on the collected data

Step 1: Generating the Key for the PRAW (Python Reddit API Wrapper) by means of an existing account on Reddit as a developer.

Step 2: Create an .ini file having the format:

[reddit_bot]

Username: reddit username

Password: reddit password

client_id: client_id that was received.

client_secret: client_secret that was received.

The username is the username used by the user to login and the password is the authentication key for the user. The client_id and the client_secret was generated by Reddit, which allows the bot to scrap its content for analysis.

Step 3: Run the Python Code to fetch the data from the past year (scrap data) and to calculate and plot the data for further analysis.

Step 4: The data set taken can be then used to find the sentiments of users generally as well as for specific topics. It can also be used to find frequency distribution of words appearing in the corpus.

Step 5: Use NLTK (The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. NLTK includes graphical demonstrations and sample data) and VaderSentiment (Valence Aware Dictionary and sEntiment Reasoner is a sentiment analysis tool that is designed to understand the emotional tone of text. It is particularly attuned to the nuances of social media text, such as tweets, but is also effective on other forms of text. VADER uses a combination of a sentiment lexicon and a set of rules that capture the intensity and polarity of sentiments. It can accurately detect positive, negative, and neutral sentiments, and it accounts for various linguistic constructs like punctuation, capitalization, degree modifiers, and conjunctions to refine its sentiment scores. This makes VADER a robust tool for sentiment analysis in natural language processing tasks) to find the positive response, negative or the neutral responses of the users and that analysis in turn depicts the nature of the topic.

4. RESULTS AND OBSERVATIONS

4.1 Choosing topics and seeing their general sentiments

The Firstly, let me define a list of words that I got from the scaping process and their classifications:

Positive comments like good, fine, happy, fun, amazing, lovely, adventurous, advocated, affability, affable, affably, affectation, affection, affectionate, affinity, affirm, affirmation, affirmative, affluence, affluent, afford, affordable, affordably, affordable, agile, agilely, agility, agreeable, agreeableness, agreeably, all-around, alluring, alluringly, altruistic, altruistically, amaze, amazed etc.

Neutral words like coarse, detached, indifferent, listless, skeptical, serious, solemn, weary etc.

Negative words like abuse, abused, abuses, abusive, abysmal, abysmally, abyss, accidental, accost, accursed, accusation, accusations, accuse, accuses, accusing, accusingly, acerbate, acerbic, stall, Stalls, stammer, stampede, standstill, stark, starkly, startle, startling, startlingly, starvation, starve, static, steal, stealing, steals, steep, steeply, stench, stereotype, stereotypical, stereotypically etc.

From the India subreddit first we see the general sentiments of the top 220 "hot" (current most popular) posts:

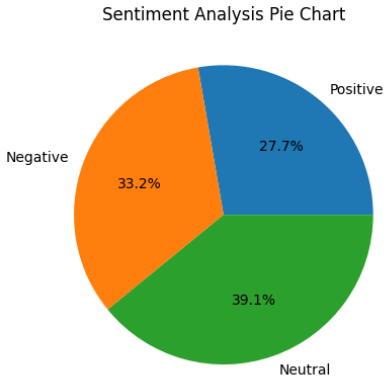

Fig 1: General Sentiment of the most popular posts on the r/India subreddit. [n=220]

Then we pick some topics that could show us a varied response throughout the subreddit. [8] One of the first topics was

"demonetization" which was probably the biggest event to happen in India in the last decade. In November 2016, the Government of India, led by Prime Minister Narendra Modi, announced the demonetization of all ₹500 and ₹1,000 banknotes. This sudden move aimed to combat black money, counterfeit currency, and corruption. The demonetized notes were to be replaced by new ₹500 and ₹2,000 notes. While the move was praised for its boldness and intention, it also faced significant criticism for the hardships it imposed on the general populace and its impact on the economy. Here is the distribution of the top posts and comments with the demonetization regex:

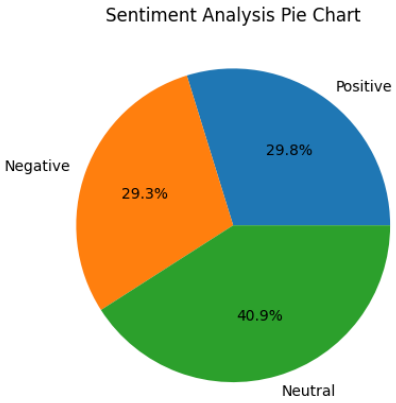

Fig 2: Sentiment of posts dealing with demonetization. [n=215]

The next topic to be analyzed is Karnataka, which is a state in India. This would show us the opinions of the general public towards Karnataka. Posts that contributed to this included a lot of news articles that contained negative-oriented headlines. The distribution is as follows:

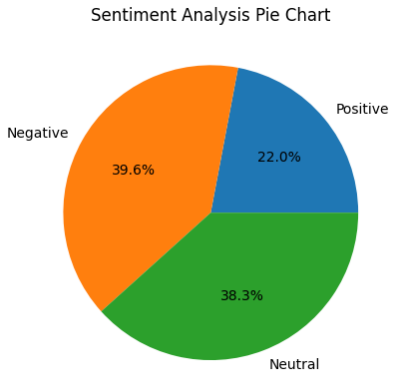

Fig 3: Sentiment of posts dealing with Karnataka. [n=227]

The final topic from the India subreddit, I wanted to check people's opinions on a popular brand and decided to use "Zomato". It is a very popular online food-ordering service (akin to UberEats) which has been in the news recently due to it launching its IPO. The distribution is as follows:

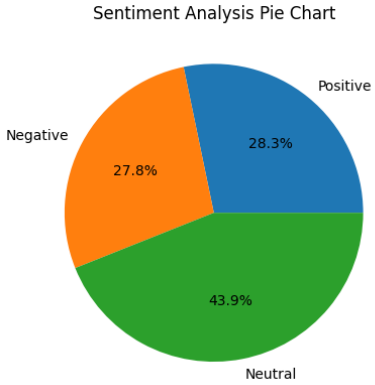

Fig 4: Sentiment of posts dealing with Zomato. [n=223]

4.2 Comparing the sentiment towards a common object across its most popular subreddits

Reddit doesn't publicly disclose detailed information about individual advertisers or how much specific companies spend on advertising on their platform. However, a study by Semrush [9] gives us the top 10 spenders in digital marketing across several industries. The data includes US-based advertising activity between January 2021 and May 2023, from 956 domains across 12 industries. I will be checking the sentiments associated with each top spender across r/all (which is a collection of the most upvoted posts across its most popular subreddits such as r/AskReddit, r/gaming, r/funny, r/worldnews, r/aww etc). [n=250]

The top 2 digital marketing spenders in the retail industry are Amazon and Target.

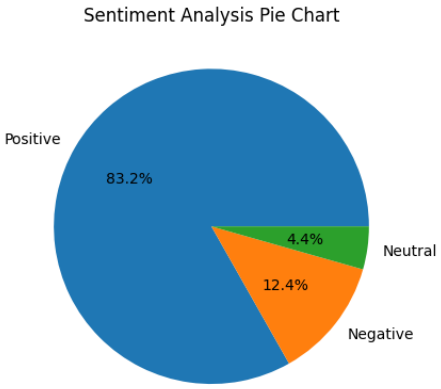

Fig 5: General sentiment for Amazon advertisements.

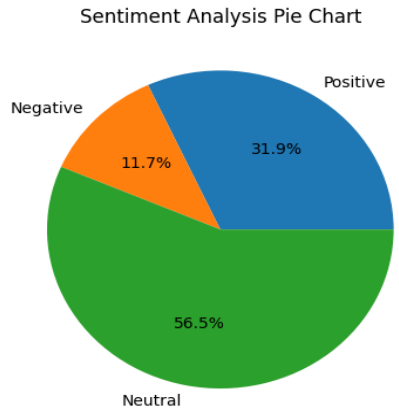

Fig 6: General sentiment for Target advertisements.

The top 2 digital marketing spenders in the SaaS industry are Adobe and Grammarly.

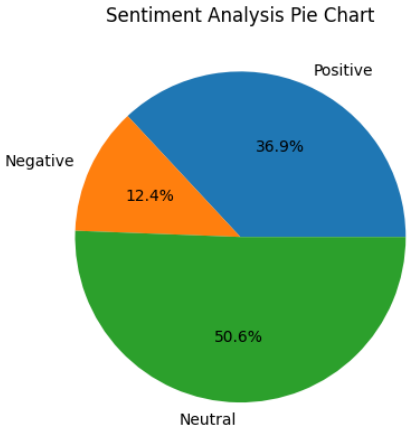

Fig 7: General sentiment for Adobe advertisements.

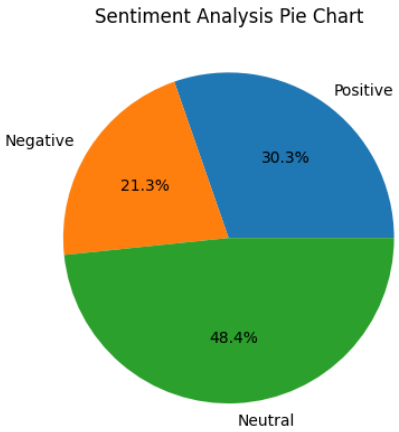

Fig 8: General sentiment for Grammarly advertisements.

The top 2 digital marketing spenders in the streaming industry are Paramount Plus and Hulu.

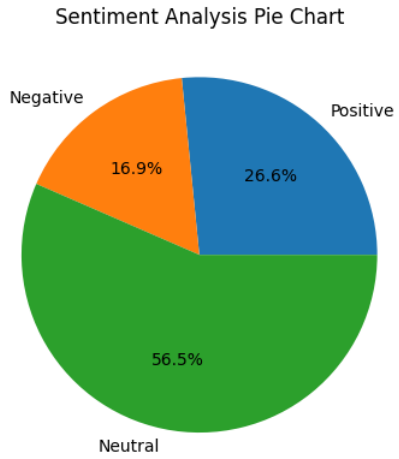

Fig 9: General sentiment for Paramount Plus advertisements.

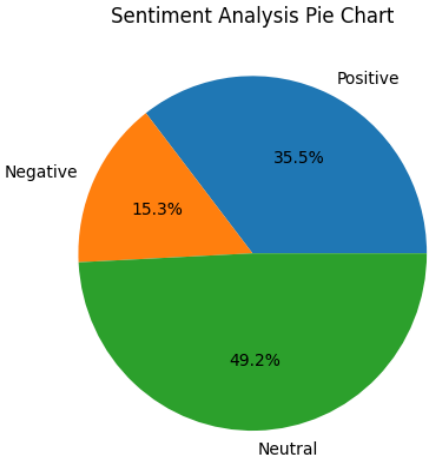

Fig 10: General sentiment for Hulu advertisements.

As we can see, there are some very blatant differences between the audiences' perception of the brands. One can conduct further nuanced research on this to see if one can skew the sentiments of the subreddits to their favor.

5. CONCLUSIONS

This paper provides us a data corpus on r/India and overview of reddit's comment structure. It also shows the relations of different brands with the general online audience. It also identified interesting relations and web platform properties among subreddits. This research also showed that utilizing NLTK and vaderSentiment for sentiment analysis on Reddit posts and comments is suitable and works fine. It can give one insights about the ongoing tends on specific topic to realize the sentiments of people. There are a few concerns still, such as subjectivity and context; sarcasm and irony; bias and misrepresentation; ethical concerns and generalization. Application of such methods can be used for marketing, evaluating guests and make operational improvements or capital expenditure.

6. REFERENCES

[1] Curiskis, S.A., Drake, B., Osborn, T.R., Kennedy, P.J.: An evaluation of document clustering and topic modelling in

two online social networks: Twitter and Reddit. Inf. Process. Manage. 57(2), 102,034 (2020)

[2] Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, p. 271. Association for Computational Linguistics (2004)

[3] Choi, D., Han, J., Chung, T., Ahn, Y.Y., Chun, B.G., Kwon, T.T.: Characterizing conversation patterns in Reddit: from the perspectives of content properties and user participation behaviors. In: Proceedings of the 2015 ACM on Conference on Online Social Networks, pp. 233–243 (2015)

[4] Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 412–418 (2004)

[5] Kouloumpis, E., Wilson, T., & Moore, J. (2021). Twitter Sentiment Analysis: The Good the Bad and the OMG!. Proceedings of the International AAAI Conference on Web and Social Media, 5(1), 538-541. https://doi.org/10.1609/icwsm.v5i1.14185

[6] Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. ICWSM 7(21), 219–222 (2007)

[7] Glenski, M., Pennycuff, C., Weninger, T.: Consumers and curators: browsing and voting patterns on Reddit. IEEE Trans. Comput. Soc. Syst. 4(4), 196–206 (2017)

[8] Stoddard, G.: Popularity and quality in social news aggregators: a study of Reddit and hacker news. In: Proceedings of the 24th International Conference on World Wide Web, pp. 815–818 (2015)

[9] Semrush Blog, accessed on 20th of April 2024, < https://www.semrush.com/blog/companies-spend-on-advertising-study/#biggest-digital-advertisers-by-industry