DATA MANAGEMENT ASSIGNMENT II (MGT7178)

COMPANY- Twitter 

APPLICATION- Sentiment Analysis

Word Count- 2719 (Including references)

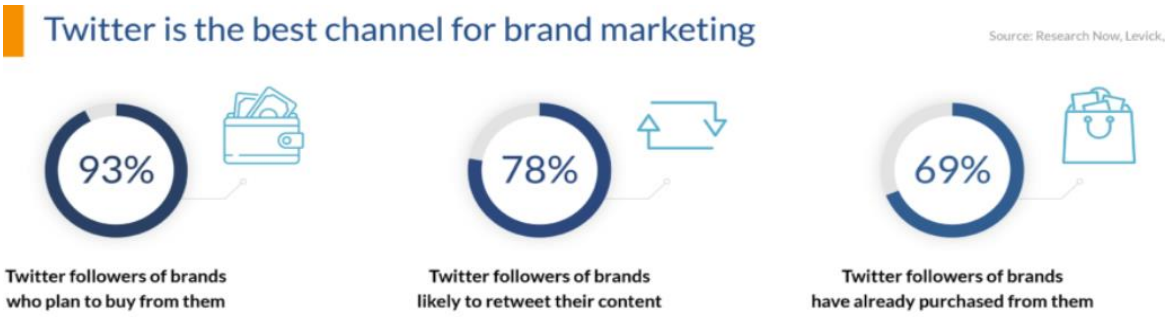
NAME- MOON KARMAKAR

ID- 40389123

Background and Context

Twitter, is a popular microblogging site with an estimated 554.7 million users worldwide who regularly use the site and send out 58 million "tweets" daily. You may also ask one of the 135,000 new members that join the network every day(Kwak et al., 2010). Twitter users either follow or are followed. When someone follows someone on Twitter, they automatically receive all their messages, or tweets, from that person. The custom of replying to tweets has developed into a distinct culture of markup: Retweet is abbreviated as RT, and "@" and "#" are used to address users. On Twitter, users may share their thoughts and opinions, advertise their products, and engage in debates with millions of other users by sending 140-character or fewer brief tweets. In fact, companies that produce these items have started to read these microblogs to determine how the general public feels about their products. These companies often read customer reviews and reply to comments on microblogs. A challenge is identifying and describing the general qualities of a feeling using technology. In this study, we outline a sentiment analysis method for Twitter data(Budiharto & Meiliana, 2018).

Twitter has been referred to as a viral marketing tool, an electronic word-of-mouth tool, and a platform for online word-of-mouth branding. It, however, differs from other marketing communications platforms, which include one-to-one (such as email), one-to-many (such as mass media), and many-to-many communication channels (e.g., the web and online groups)(Hoffman & Novak, 1996). Organizations utilize sentiment analysis or opinion-mining techniques to examine user feedback. Sentiment analysis is defined as “the task of finding the opinions of authors about specific entities.” Based on and evaluated at the document, phrase, or word level are sentiment analyses. For Twitter, we base our analysis on the entire tweet and presume that it comprises an opinion or a feeling (ACM & 2013, 2013).



Percentage of Brands Promoted on Twitter (Market Analysis of Twitter, n.d.)

Analysis of large-scale social data, which confirms corporate perspective from customer perspectives, is especially critical for top-level management to help resolve the practical problem. These customer opinions can affect how consumers see a brand and how devoted and ardent they are to it. With social media monitoring and sentiment analysis, businesses may gain access to consumer insights to enhance product quality, offer better customer service, or even spot new business opportunities (Oliveira et al., 2014).

Twitter Data Sentiment Analysis using Hadoop

Although there is an increasing amount of interest in Twitter data which can be properly gathered and analyzed data using Twitter APIs. Users must construct, send, and receive queries, manage rate limits, and wrangle nested and real-time feedback objects into analysis-friendly data models in addition to learning and ingesting the necessary information from Twitter's developer documentation in order to interact with the company's APIs. Fortunately, the retweet R package is made to make these procedures simple, increasing the accessibility of Twitter's APIs to a larger range of user types (software & 2019, 2019).

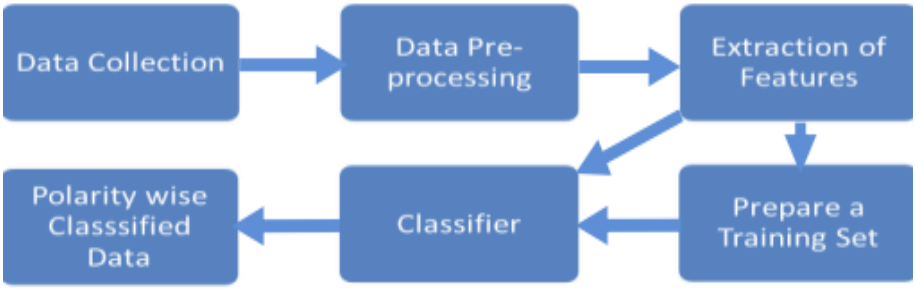
Hadoop has been used by some academics to analyze the sentiment of Twitter data. They classified the many tweets using a Naive Bayes algorithm. Utilizing Apache Flume and Apache Pig, other researchers have suggested an effective method for doing opinion mining on data from Twitter. Some have also proposed a system that uses Hadoop Mahout Algorithms to analyze public opinion and uses Twitter data to offer feedback by executing sentiment analysis using the Hadoop framework, where Hadoop is utilized to extract information with the aid of Cloudera setup. Using the Twitter4j software, which internally makes use of the Twitter REST API, tweets were gathered. (Lee et al., 2012).

In addition to tweets, Twitter data is supplied into Apache Pig in a highly nested JSON format and may also include photos, URLs, user ids from Twitter, tweet ids, user profiles, the location from where the tweets were posted, the time the tweets were written, and other details. The sentiment analysis just takes into account tweets. As a pre-processing stage for sentiment analysis, we removed the Twitter id and tweet metadata from the JSON Twitter data. (Ingle et al., n.d.).

How Sentiment Analysis Works in Twitter Data

The mining of Twitter data is a difficult process. The information gathered is unprocessed information. It is crucial to pre-process or clean the raw data before using a classifier. Uniform casing, deletion of emoticons, URLs, stop words, slang word decompression, and elongation word compression is all part of the pre-processing process. The pre-processing process is demonstrated in the stages that follow (Desai & Mehta, 2017)

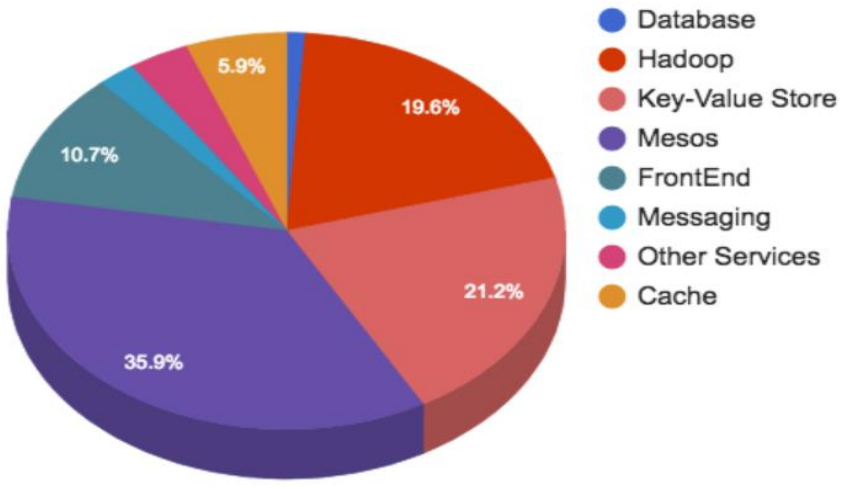
- Remove any Twitter symbols like the account ID (@), hashtags (#), and retweets (RT).
- Take out the emoticons, hyperlinks, and URLs. Given that we are only working with text data, it is vital to eliminate non-letter data and symbols.
- Eliminate stop words like "are," "is," "am," etc. The stop words are meant to exclude any emotional content from sentences in order to reduce dataset size.
- Shorten words with extra letters, like happyyy, to happy.
- G8, F9, and other slag words should be compressed. Slang words typically function as nouns or adjectives and express extremely strong emotions. Therefore, they must be decompressed (Desai & Mehta, 2017).



The Sentiment Analysis Process of Twitter Data(*IEEE Xplore Full-Text PDF:*, n.d.)

Technical Infrastructure of the Twitter

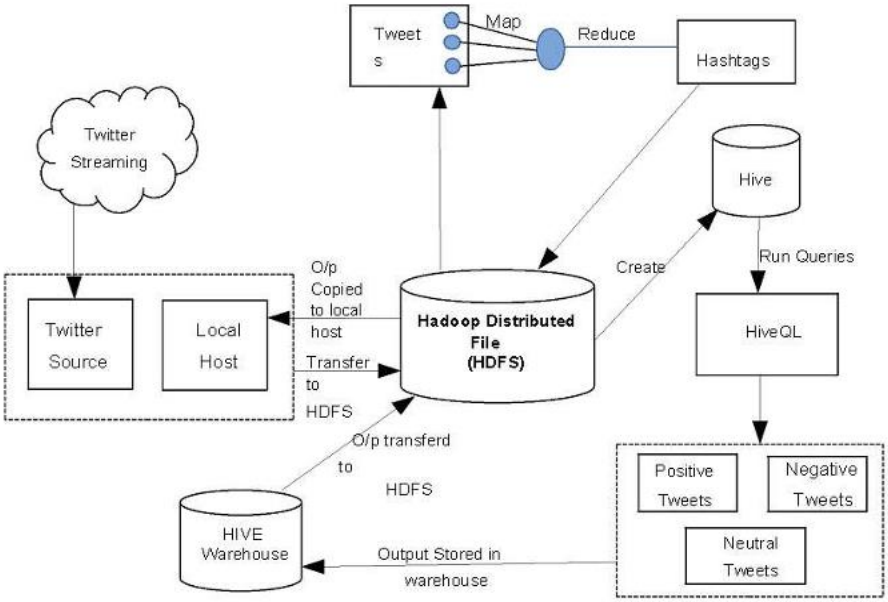
The scale, complexity, user base, and range of use cases of Twitter's analytics platform have all grown significantly in the last several years. Four persons made up the analytics team in 2010, which had a total of about 100 workers and was the only group to regularly use our 30-node Hadoop cluster. The firm now employs over a thousand people. Thousands of Hadoop nodes are spread across several data centers. Our primary Hadoop data warehouse receives roughly 100 terabytes of raw data each day, and engineers and data scientists from dozens of teams work together to perform tens of thousands of Hadoop jobs. These tasks include data cleansing, basic aggregations, and report preparation, as well as the development of data-powered products and the instruction of machine-learned models for promoted goods, spam detection, and follower suggestion(ACM & 2013, 2013).



Data used in the application

Direct database access is a highly popular design that may even come naturally when the primary application is already dependent on an RDBMS. If the database can keep up, it is an excellent

design since there are no additional systems to worry about, a tonne of tools to query the data and always-current logs. The design is deceptively straightforward: just develop a flexible schema, like the one displayed below, for instance. However, using a database for logging at scale soon fails for a variety of reasons. Twitter makes use of the Scribe system, which is reliable, fault-tolerant, and distributed for collecting large amounts of streaming log data(LinJimmy & RyaboyDmitriy, 2013).



System Architecture(Ingle et al., n.d.).

The tweets are divided into tokens, and each token has a polarity value—a floating point integer ranging from 1 to -1 assigned to it-

A. Positive Tweets: Tweets that express satisfaction or approval with something are referred to as positive tweets. Tweets like "It was an inspirational movie!!!" or "Best movie ever," for instance.

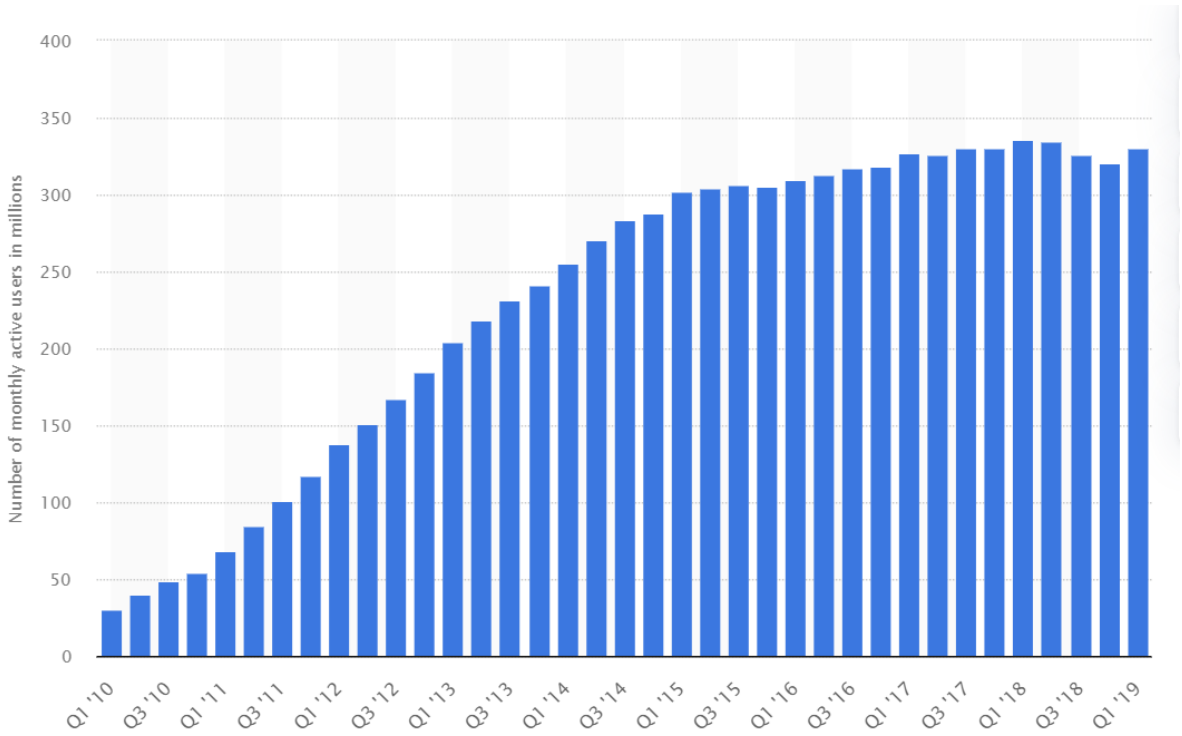
B. Negative Tweets: Tweets that express opposition or a negative reaction to anything are categorized as negative tweets. Tweets like "Waste of time" or "Worst movie ever," for instance.

C. Neutral Tweets: Neutral tweets are those that neither express support for or appreciation of anything nor do they express opposition to or denigration of it. Additionally, there are tweets that present hypotheses or facts. for instance, tweets like Earth is round(Ingle et al., n.d.).

Current and Estimated Benefits of Twitter to the Organization

Organizations may benefit from examining these user insights in order to deliver better services, increase user experience, develop product designs, and manage company success. Online communities usually allow user comments, and in the case of Twitter, the remarks can be posted in real-time or very close to it and quickly reach a large audience (Poria et al., n.d.).

However, for businesses to benefit from Twitter data, they must gather, store, and evaluate the massive amounts of data that Twitter generates every day. More than 319 million active users sent more than 500 million tweets daily in 2016.



Global monthly active Twitter user count from the first quarter of 2010 to the first quarter of 2019 (in millions) (Source- https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/)

Several of the busiest Twitter accounts get tens of thousands of messages each day (e.g., Xbox Support has more than 400,000 followers and receives more than 1.5 million tweets daily; Justin Bieber receives more than 300,000 tweets daily). There were more than 672 million tweets on the World Cup in 2014 overall (Rogers, S.,2014).
Lexalytics, Converseon, and Summize are a few examples of for-profit sentiment services that provide software to process datasets of this magnitude.

Limitations and Management of the Solution

Modern TSA (Twitter Sentiment Analysis) methods, however, are unable to provide applications with satisfactory performance. According to a survey of the literature, these applications' accuracy ranges from 40% to 70% (Abbasi et al., 2014.),(Ghiassi et al., 2016). A stronger and more precise collection of TSA tools is needed if TSA is to help firms analyze their consumer input more effectively (Ghiassi, Zimbra, & Lee, 2016). There are a number of characteristics of tweets that make TSA extremely difficult, which may be the cause of the poor performance numbers.

The vocabulary used in tweets is varied and always changing, and emojis, slang, and acronyms are frequently used. Due to the limited number of phrases available in tweets to be evaluated using a sentiment lexicon, tweet feature representations are poorly filled. Sparsity is a flaw in conventional feature representations that often reduces the effectiveness of sentiment analysis techniques (Ghiassi & Lee, 2018).

The study shows how developing and employing a language set tailored to Twitter might improve TSA accuracy. These accuracy levels can be further elevated when combined with DAN2, a machine learning technique. The issue's complexity is decreased and the feature matrix's density is increased by using a TSA-specific reduced Twitter lexicon set, both of which aid in resolving the issue of feature sparsity (Saif et al., 2012).

Lexicon-based, machine-learning and hybrid solutions are all available for TSA problems:

The lexicon-based sentiment analysis makes the assumption that each word has a previous polarity that is distinct from context. This method produces a huge number of features and will produce a relatively sparse representation of text for sentence-level texts like microblogs. A tweet will frequently just contain a small number of the several thousand elements that make up its vocabulary(Oliveira et al., 2014). (Moreno-Ortiz et al., 2013) evaluate the use of a lexicon-based methodology for sentiment analysis of Spanish-language tweets. They state that it is now a truism in sentiment analysis that building systems specially tailored for a certain topic area is a major factor in achieving effective results.

The classifiers used in the machine learning technique are typically taught using a collection of features made up of n-grams. This advantage frequently applies just to one area and is not typically transferrable to other applications. One of the reasons why researchers have developed "hybrid techniques" is due to this constraint. In cross-domain scenarios, a lexicon-based system could outperform pure or hybrid machine-learning approaches (Kennedy et al., 2006).

Machine learning is used in the majority of TSA research. Its versatility and accuracy are the reasons for its success. The labeling (scoring) of the datasets, which is frequently done manually and may need domain specialists, is a substantial work in this technique. Although it is now simpler to gather a sizable dataset for sentiment analysis, categorizing these recordings is still difficult (Pak et al., n.d.).

Conclusion

Sentiment analysis is used in a thorough approach to categorize Twitter's extremely unstructured data into positive and negative categories. Second, we have covered a variety of methods for doing knowledge-based and machine-learning methods for sentiment analysis on Twitter data(IEEE Xplore Full-Text PDF, n.d.). The method described in this study employs a hierarchical approach and starts with the collection of sizable corpora that are manually labeled in order to generate sizable training and testing datasets (Pak & Paroubek, 2010).
The process of gathering data will introduce us to the Java Twitter Streaming API. We will get the opportunity to work with Hadoop, a well-known parallel data processing platform.
Our research has a few limitations and potential directions. Our Twitter data sets are constrained in size, albeit they are thousands strong and big enough to train a machine-learned classifier. Our brand-related Twitter data sets can only be as large as the manual annotation of tweets required to produce the gold standard sentiment class labels for classifier training and assessment(Arias et al., 2013).

References

Abbasi, A., Hassan, A., Ninth, M. D.-P. of the, & 2014, undefined. (n.d.). Benchmarking twitter sentiment analysis tools. *Aclanthology.Org*. Retrieved December 20, 2022, from https://aclanthology.org/L14-1406/

ACM, R. F.-C. of the, & 2013, undefined. (2013). Techniques and applications for sentiment analysis. *Dl.Acm.Org*, *56*(4), 82–89. https://doi.org/10.1145/2436256.2436274

Arias, M., Arratia, A., Intelligent, R. X.-A. T. on, & 2014, undefined. (2013). Forecasting with twitter data. *Dl.Acm.Org*, *5*(8). https://doi.org/10.1145/2542182.2542190

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big Data*, *5*(1). https://doi.org/10.1186/S40537-018-0164-1

Desai, M., & Mehta, M. A. (2017). Techniques for sentiment analysis of Twitter data: A comprehensive survey. *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2016*, 149–154. https://doi.org/10.1109/CCAA.2016.7813707

Ghiassi, M., & Lee, S. (2018). A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications*, *106*, 197–216. https://doi.org/10.1016/J.ESWA.2018.04.006

Ghiassi, M., Zimbra, D., Information, S. L.-J. of M., & 2016, undefined. (2016). Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Taylor & Francis*, *33*(4), 1034–1058. https://doi.org/10.1080/07421222.2016.1267526

Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks. *Journal of Management Information Systems*, *33*(4), 1034–1058. https://doi.org/10.1080/07421222.2016.1267526

Hoffman, D. L., & Novak, T. P. (1996). Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations. *Journal of Marketing*, *60*(3), 50–68. https://doi.org/10.1177/002224299606000304

IEEE Xplore Full-Text PDF: (n.d.). Retrieved December 20, 2022, from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7813707&casa_token=g0xizehY2dAAAAA:24rD69DiwlKEfVbsuuaoTHFRX-R8-SJPR0Gr19pQvo3sEELUpNKCJn-PdG_qlMippTUKcOa0Sek

Ingle, A., Kante, A., Samak, S., of, A. K.-I. J., & 2015, undefined. (n.d.). Sentiment analysis of twitter data using hadoop. *Pnrsolution.Org*, *3*(6). Retrieved December 21, 2022, from http://www.pnrsolution.org/Datacenter/Vol3/Issue6/18.pdf

Kennedy, A., intelligence, D. I.-C., & 2006, undefined. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Wiley Online Library*, *22*(2), 110–125. https://doi.org/10.1111/j.1467-8640.2006.00277.x

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 591–600. https://doi.org/10.1145/1772690.1772751

Lee, G., Lin, J., Liu, C., Lorek, A., & Ryaboy, D. (2012). The unified logging infrastructure for data analytics at twitter. *Proceedings of the VLDB Endowment*, *5*(12), 1771–1780. https://doi.org/10.14778/2367502.2367516

LinJimmy, & RyaboyDmitriy. (2013). Scaling big data mining infrastructure. *ACM SIGKDD Explorations Newsletter*, *14*(2), 6–19. https://doi.org/10.1145/2481244.2481247

Market analysis of twitter - Google Search. (n.d.). Retrieved December 20, 2022, from https://www.google.com/search?q=Market+analysis+of+twitter&rlz=1C1CHBF_enIN917IN917&sxsrf=ALiCzsZOPijzM-IXas0aho4-fzKVn6lWWg%3A1671453550485&ei=blugY66WHdCegQaJyInACw&ved=0ahUKEwiuhYrZ2YX8AhVQT8AKHQlkArgQ4dUDCBA&uact=5&oq=Market+analysis+of+twitter&gs_lcp=Cgxnd3Mtd2l6LXNlcnAQAzIFCAAQogQyBQgAEKIEMgUIABCiBDoHCCMQsAMQJzoKCAAQRxDWBBCwA0oECEEYAEoECEYYAFDMBFiqFWDWGGgBcAB4AIABdIgB0QGSAQMxLjGYAQCgAQHIAQnAAQE&sclient=gws-wiz-serp

Moreno-Ortiz, A., lenguaje, C. H.-P. del, & 2013, undefined. (2013). Lexicon-based sentiment analysis of Twitter messages in Spanish. *Journal.Sepln.Org*. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4664

Oliveira, N., Cortez, P., international, N. A.-P. of the 18th, & 2014, undefined. (2014). Automatic creation of stock market lexicons for sentiment analysis using stocktwits data. *Dl.Acm.Org*, 115–123. https://doi.org/10.1145/2628194.2628235

Pak, A., … P. P. the S. I. C. on, & 2010, undefined. (n.d.). Twitter as a corpus for sentiment analysis and opinion mining. *Aclanthology.Org*. Retrieved December 20, 2022, from https://aclanthology.org/L10-1263/

Pak, A., & Paroubek, P. (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

Poria, S., Cambria, E., Systems, A. G.-K.-B., & 2016, undefined. (n.d.). Aspect extraction for opinion mining with a deep convolutional neural network. *Elsevier*. Retrieved December 20, 2022, from https://www.sciencedirect.com/science/article/pii/S0950705116301721?casa_token=jFnvUf5DTu4AAAAA:oDb6oUVn59f14s-QDjfMhIqNBi5GuxJ4O2VNcpjP4a8ZFPXbHVaXnDQ0P0Ox2IaMc7vSZlRj280

Rogers, S. . Insights into the WorldCup conversation... - Google Scholar. (n.d.). Retrieved December 20, 2022, from https://scholar.google.com/scholar?q=Rogers,%20S.%20.%20Insights%20into%20the%20WorldCup%20conversation%20on%20Twitter.%20https:blog.twitter.com2014insights-into-the-worldcup-conversation-on-twitter%20Accessed:%202014.08.15.

Saif, H., He, Y., & Alani, H. (2012). *Alleviating data sparsity for twitter sentiment analysis*. 2–9. http://oro.open.ac.uk/38501/

software, M. K.-J. of open source, & 2019, undefined. (2019). rtweet: Collecting and analyzing Twitter data. *Joss.Theoj.Org*. https://doi.org/10.21105/joss.01829