Global Scientific JOURNALS

GSJ: Volume 6, Issue 2, February 2018, Online: ISSN 2320-9186

www.globalscientificjournal.com

# STUDENT'S SENTIMENTS ON FACEBOOK: AN ANALYSIS USING BIG DATA ANALYTICS AND DATA MINING TECHNIQUES

GEORGE M. GRANADOS,RENATO DAN A. PABLO II

## ABSTRACT

Data mining techniques such as decision trees, classification, and clustering can be used to solve the problem of Big Data. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).Facebook is a famous social media application that connects people around the globe.MCC Files is a Facebook group that allows users, mostly Mabalacat City College student to post anything and everything they wanted to. The purpose of the study is about the development of an algorithm that can compute and compare words in the MCC Files group posts and the AFINN database. The researches managed to download MCC Files posts from July 03, 2015 until July 15, 2017 as the primary data set using Facepager. There were originally 13852 posts in the data set but after the data mining process has been used, only 4783 unique posts remained. The remaining unique posts were cleansed using Data Cleaner and MS Excel. The overall sentiments of MCC Files are generally positive based on the experimental sentiment scoring method used in the study.

Keywords: Data Mining, Big Data Analytics, Sentiments Analysis

## INTRODUCTION

Data mining techniques such as decision trees, classification, and clustering can be used to solve the problem of Big Data. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

The key properties of data mining are (1) Automatic discovery of pattern, (2) Prediction of likely outcomes, (3) Creation of actionable information, and (4)Focus on large data sets and databases

Data mining can answer questions that cannot be addressed through simple query and reporting techniques.



Facebook is a famous social media application that connects people around the globe. According to the Zephoria statistics, as of December 2016, there are about 1.79 Billion active users in Facebook (Zephoria.com, 2016). Most common use of Facebook is to contact family members and friends. But, Facebook has been known to be a platform for sentiment postings. Different Facebook groups allow people to post their sentiments on different topics or issues. One Facebook group that serves the said purpose is the MCC Files (LASolidaridadMCC). MCC Files is a Facebook group that allows users, mostly Mabalacat City College student to post anything and everything they wanted to.

There are other Facebook groups intended for Mabalacat City College, but by far, MCC Files is the most active and widely used Facebook group by the students.

With the advent of the different social media applications such as Facebook, Instagram, and Twitter; Big Data phenomenon happened. Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate (Kundi and Asghar, 2014). Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization and querying, and information privacy. Because of this phenomenon, it is impossible if not difficult to analyze and verify the contents of the posts made to MCC Files.

The purpose of this study is to identify the polarity of sentiments of the students of Mabalacat City College through big data analytics and data mining techniques by using the experimental method of scoring students' posts in the MCC Files Facebook group.

**Literature Review**

Data Mining: Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost. Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes (1) operational or transactional data such as, sales, cost, inventory, payroll, and accounting, (2) nonoperational data, such as industry sales, forecast data, and macro-economic data, (3) meta data - data about the data itself, such as logical database design or data dictionary definitions(Zentut.com, 2017).

The patterns, associations, or relationships among all this *data* can provide *information*.

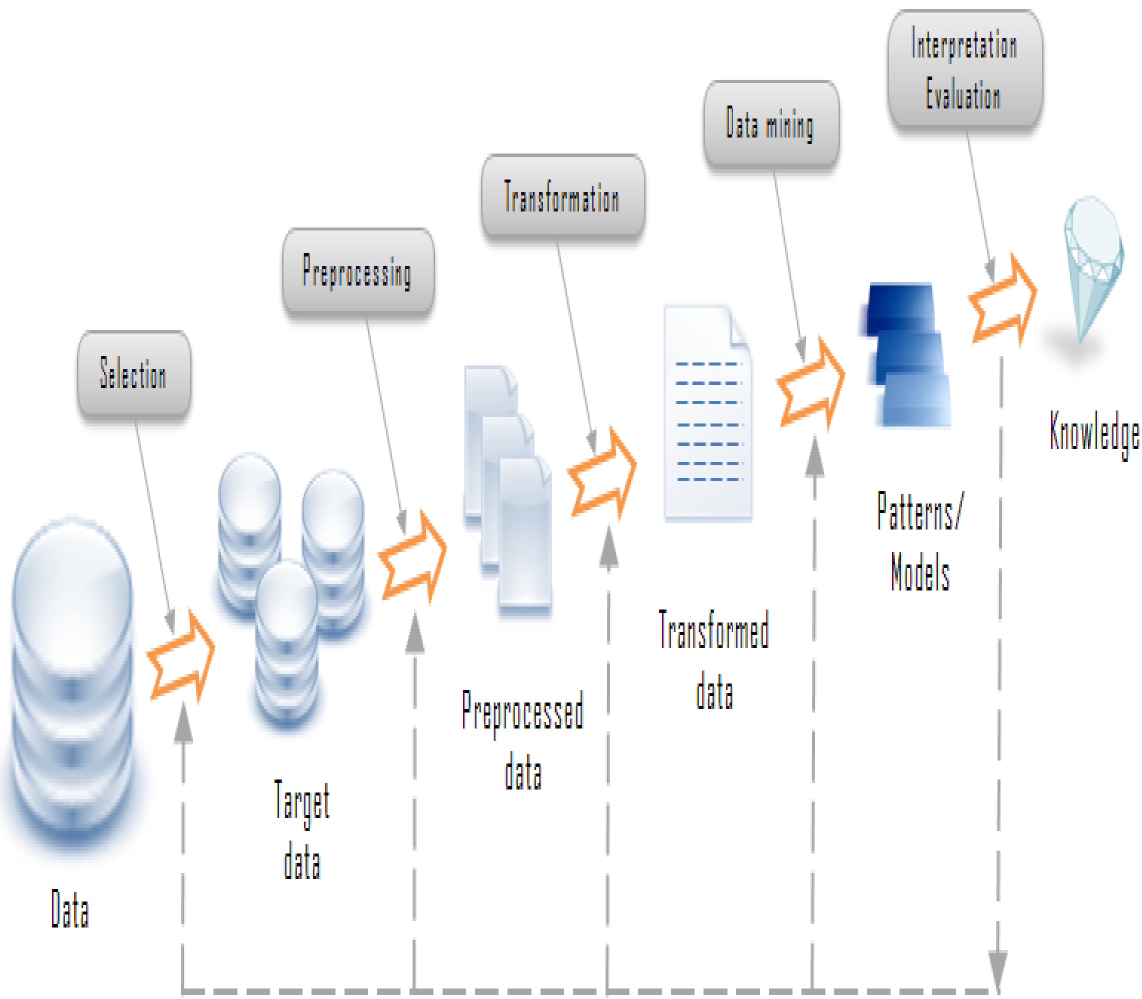


Figure 1: Data Mining Process
*(Source: http://www.zentut.com/data-mining/what-is-data-mining/)*

For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining (Hu and Liu, 2012).

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

WalMart is pioneering massive data mining to transform its supplier relationships. WalMart captures point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits this data to its massive 7.5 terabyte data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, WalMart computers processed over 1 million complex data queries (LoyalRewards.com, n.d.).

The National Basketball Association (NBA) is exploring a data mining application that can be used in conjunction with image recordings of basketball games. The Advanced Scout software analyzes the movements of players to help coaches orchestrate plays and strategies. For example, an analysis of the play-by-play sheet of the game played between the New York Knicks and the Cleveland Cavaliers on January 6, 1995 reveals that when Mark Price played the Guard position, John Williams attempted four jump shots and made each one! Advanced Scout not only finds this pattern, but explains that it is interesting because it differs considerably from the average shooting percentage of 49.30% for the Cavaliers during that game.

By using the NBA universal clock, a coach can automatically bring up the video clips showing each of the jump shots attempted by Williams with Price on the floor, without needing to comb through hours of video footage. Those clips show a very successful pick-and-roll play in which Price draws the Knick's defense and then finds Williams for an open jump shot.

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks (Asli, Hakkani-tur, and Feng, 2010). Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.



Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a

given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.
- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships (Zhou, 2015).

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. NCR has the capacity to deliver applications exceeding 100 terabytes. There are two critical technological drivers:

- Size of the database: the more data being processed and maintained, the more powerful the system required.
- Query complexity: the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

Big Data Analytics: Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things (Olson and Delen, 2008).

Semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems.

Sentiments Analysis: Sentiment analysis – otherwise known as opinion mining – is a much bandied about but often misunderstood term. In essence, it is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention (Bannister, 2015). Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Sentiment analysis is a branch of natural language processing or machine learning methods. It becomes one of the most important sources in decision making (El Din, 2016). Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to obtain researcher's feelings expressed in positive or negative reviews or opinion by analyzing a big number of documents and papers (Wem, Dai, and Zhao, 2011).

AFINN Database: AFINN database is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011 (Nielsen, 2011).

There are two original versions of the AFINN database:

1. AFINN-111 Version: 2477 words and phrases.
2. AFINN-96: 1468 unique words and phrases on 1480 lines.

The latest and newest version of the AFINN database is AFINN 164 which contains 3478 positive and negative words and emoticons.

The AFINN database was used in the study "Good Friends, Bad News - Affect and Virality in Twitter" (Nielsen et. al, 2011) which tried to measure the polarity and sentiments of various Twitter pages and posts using the AFINN database.

Currently, a new AFINN database is in the works that will include three (3) or more common word phrases that are found in social media such as Twitter, Facebook, Youtube and others. The word phrases will also be scored just like in the previous AFINN databases (+1 to -1) that aims to extract a more accurate sentiments of posters in these social media.

**Conceptual Framework**

The study's conceptual framework is shown in figure 2. The input data came from MCC Files; the AFINN database the data will be processed using data mining techniques, decision trees and the scoring methods. The output will be the sentiment score of a post. The analysis of the data retrieved and processed will follow.

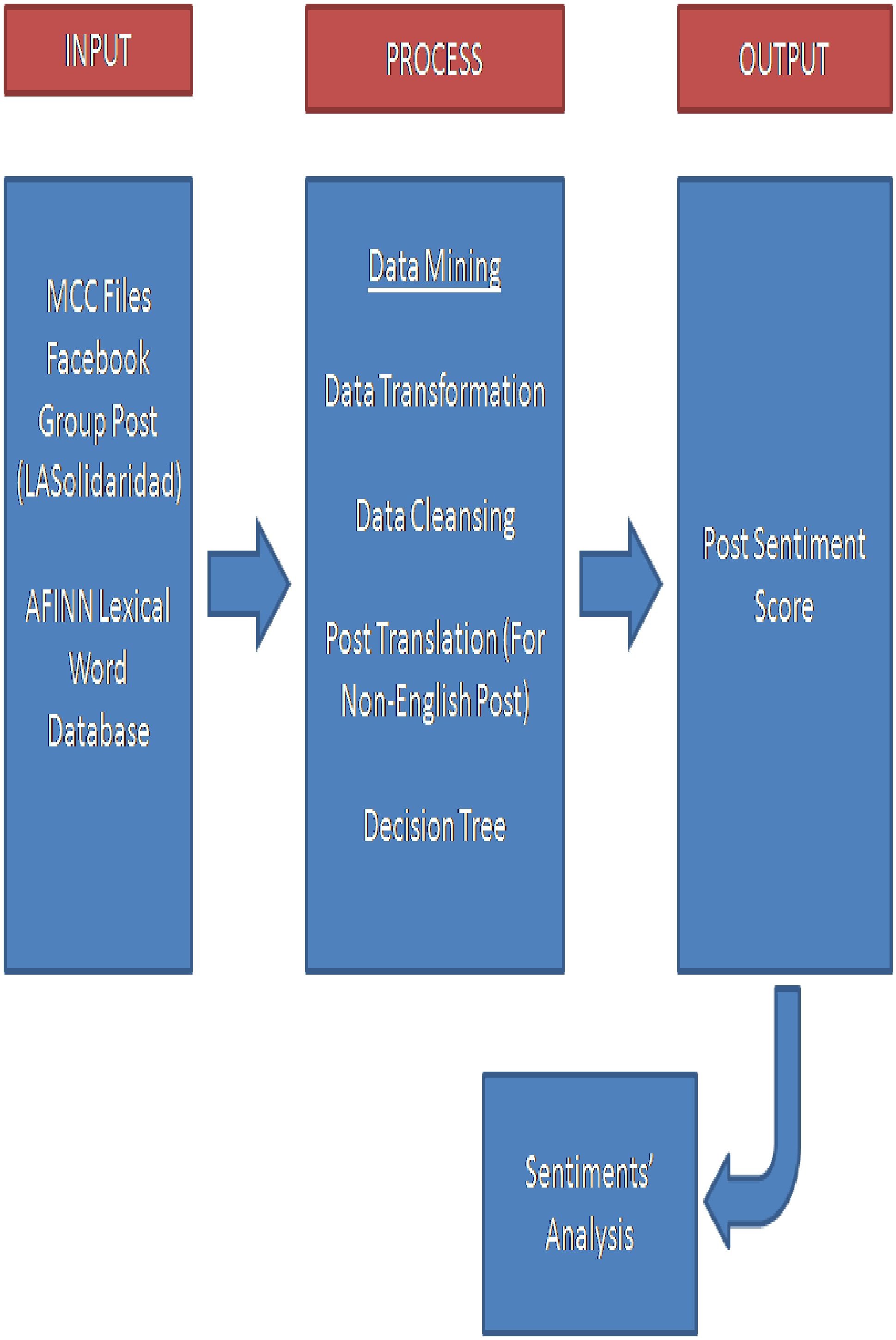


Figure 2: Conceptual Framework

## METHODOLOGY

The study is about the development of an algorithm that can compute and compare words in the MCC Files group posts and the AFINN database. The focus of the research was to identify the general sentiments of the users posting in MCC Files (LASolidaridad) using the experimental sentiment scoring method.

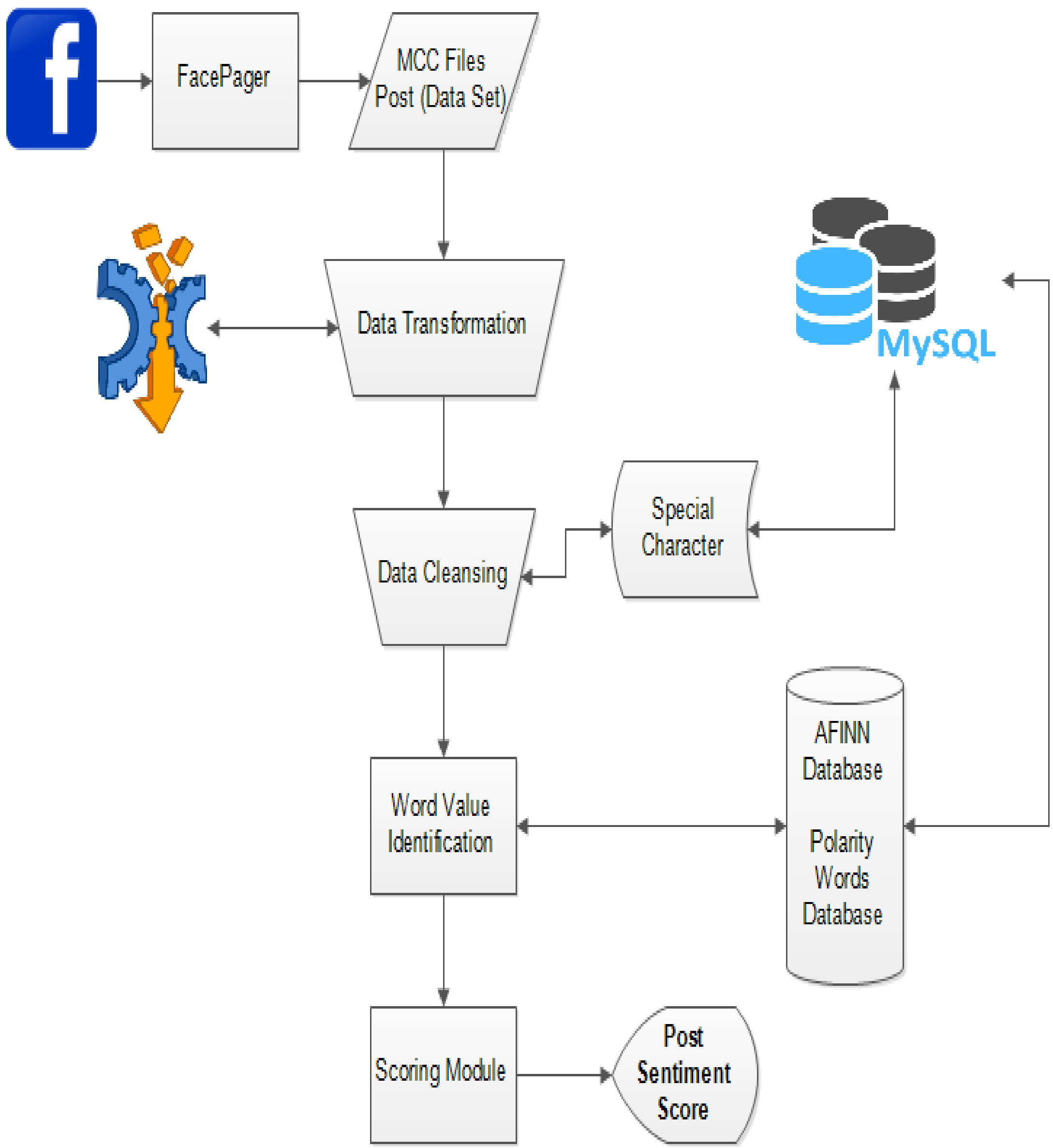


Figure 3: Sentiments Analysis Data Flow

Shown in figure 3 is the methodology of the study. Facepager was used to download the posts made to the MCC Files group. The downloaded posts were processed for data mining. The downloaded files were submitted to Data Transformation. Data transformation changed the format of the downloaded files that is compatible to the data cleansing tool and MySQL database. The downloaded posts were submitted for data cleansing or data scrubbing. Data cleansing detected and removed unnecessary words in the downloaded posts such as articles, special characters, and neutral words. Word value identification followed the data mining process. The scoring module used the experimental sentiment scoring method.

The experimental sentiment scoring method would be the sentiment scoring using the AFINN 164 database. This step checked if the words in the post are saved in the AFINN database. If a match is retrieved, the score of the sentiments word is also retrieved and added to the total. The final total score will be the sentiment score of the post shown in figure 3.

**Word Comparison Algorithm**

The formula for comparing the posts that was used in the experimental scoring methods of the study is:

*percent = (1 - (lev /max(strlen(WORD),strlen(AFINN)) ) ) *100;*

Where:

WORD = the word in a given post

AFINN = the word saved in the AFINN database

percent = percentage score of the WORD and AFINN comparison

lev = the distance of the words in comparison using Levenshtein algorithm

**Experimental Sentiment ScoringMethod**

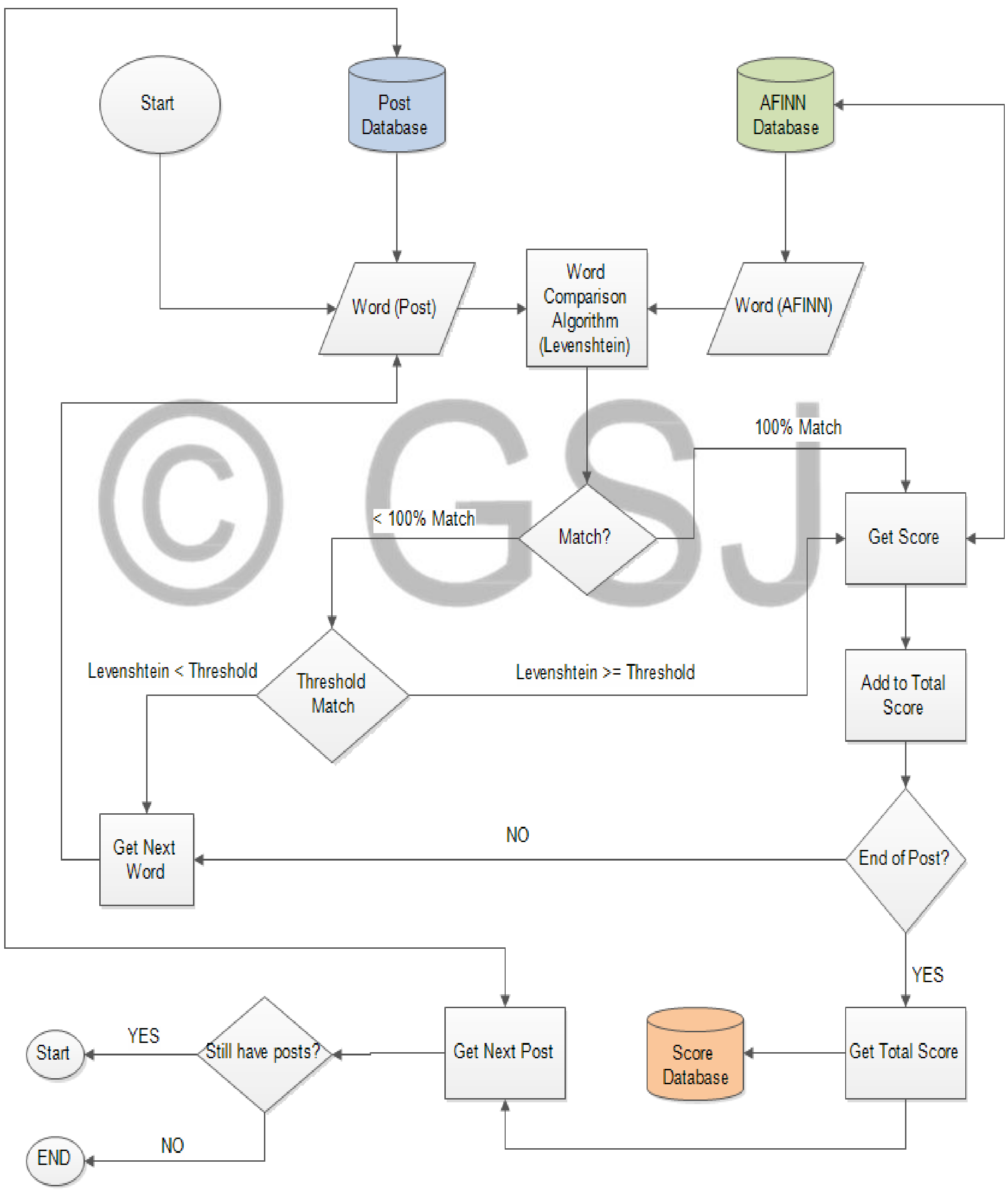


Figure 4: Decision Tree for the Experimental Scoring Method

The scoring of the posts based on the AFINN database was derived in a decision tree that identifies if WORD and AFINN is a complete match (100%) or above the comparison threshold.

In this study, the experimental sentiment scoring method was to calculate the sentiment scores of the posts made in MCC Files. Further, the top 200 posts in terms of word count will be used as test data sets for the experimental scoring method. The posts were scored on its original form (un-translated) and in the translated form. The translation was done using the Google Translate API (translate.google.com). Google Translate is the most accurate word translator in the web today.Various data set collections were also be used to test the experimental sentiment scoring method. The data set came from different websites such as Amazon, Twitter, and NexTel, among others. The data set is composed of 8533 sentiments on various brands and circumstances.

## RESULTS

The researches managed to download MCC Files posts from July 03, 2015 until July 15, 2017 as the primary data set using Facepager. There were originally 13852 posts in the data setbut after the data mining process has been used, only 4783 unique posts remained. The remaining unique posts were cleansed using Data Cleaner and MS Excel.

The processing and application of the data mining process in the data set took time to finish considering the amount of data that the tools needed to process. It was recorded that the average words per post in the data set is 425. The data set was processed for a word count of each individual post to determine the top 200 post that was used as the data set for the study. After the first data set was used, the data was submitted for a translation from "TAGLISH" to full English using the Google Translate API. The whole data set was not translated because some of the posts were written in "Kapampangan" which is not supported by Google Translate API.

| Breattaking (WORD) | Breathtaking (AFINN) |
|---|---|

(1) Lev = levenshtein (WORD,AFINN) : answer is 1

(2) *percent = (1 - (lev /max(strlen(WORD),strlen(AFINN)) ) ) *100;*

(3) *percent = (1 - (lev /max(strlen(*Breattaking *),strlen(*Breathtaking*)) ) ) *100;*

(4) *percent = (1 - (lev /max(11,12)) ) ) *100;*

(5) *percent = (1 - (1/max(11,12)) ) ) *100;*

(6) *percent = 92*

Figure 5: Sample Word Comparison Calculation

Shown in figure 5 is the Word Comparison Algorithm used in the study was proven accurate. A threshold of 90% was used for the data sets. This means that WORD and AFINN must match at least 90% (threshold) to be considered as the same word. Less than the declared threshold, the word will not be scored from the AFINN database.

The results also showed that the algorithm is more accurate to words with more than four (4) characters.

The experimental sentiment scoring method was used with the following data sets:

A. Data Set 1: Top 200 (word count) MCC Files Original Posts(un-translated)

B. Data Set 2: Top 200 (word count) MCC Files Translated Posts

C. Data Set 3: 8533sentiments from various sources.

Data Set 1 produced an overall sentiment score of 397 points with an average of 1.99 sentiment score per post (See Appendix A). The sentiment score indicating positive sentiments among the majority of the users.

Data Set 2 produce and overall score of 953.8 points with an average of 4.77 sentiment score per post (See Appendix B). The sentiment score also indicated a positive sentiment among the majority of the users.

Table 1 shows the difference of Data Set 1 and Data Set 2 in terms of overall sentiment score and average sentiment score per post.

Data Set 1 and Data Set 2 produced two different results because of the translations made from "FILIPINO - TAGLISH" to English in Data Set 1 using the GOOGLE TRANSLATE API. Translated posts provided more words which can be scored by the experimental sentiment scoring method using the AFINN database. The experimental sentiment scoring method showed an accurate result in Data Set 2 more than Data Set 1 because of the added words from the original after Data Set 1 was translated (See Appendix C for Data Sets 1 and 2 Comparison).

**Table 1: Comparison of Data Set 1 and Data Set 2**

| Data Set | Overall Sentiment Score | Average Sentiment Score |
|---|---|---|
| Data Set 1 | 397 | 1.99 |
| Data Set 2 | 953.8 | 4.77 |

Data Set 3 produced a sentiment score of 688 points, with an average of 0.08 sentiment score per review as shown in table 2.

It is shown that the average sentiment scoreis somewhat low, this is because of the number of reviews that were processed. And the reviews were just one sentence long, there were only a handful of reviews that were more than one (1) sentence compared to Data Sets 1 and 2.

**Table 2: Data Set 3 Sentiment Score**

| Data Set | Overall Sentiment Score | Average Sentiment Score |
|---|---|---|
| Data Set 3 | 688 | 0.08 |

The experimental sentiment scoring method showed accuracy in determining the polarity of each review based on the AFINN database as shown in Table 3 for the Top 10 Data Set 3 sentiment score and Table 4 for the Last 10 sentiment score (See Appendix E for the Top 100 Data Set 3 Sentiment Score and Appendix F for the Last 100 Data Set 3 Sentiment Score).

**Table 3: Top 10 Data Set 3 Sentiment Score**

| No | Review ID | Sentiment | Score |
|---|---|---|---|
| 1 | 4840 | I have always liked and admired Hillary Clinton, she has a fine soul, great brain and amazing life accomplishment list. | **3.2** |
| 2 | 31206 | Anyone that thinks open seating on Southwest Airlines is a great idea obviously doesn't know what 's like to be indecisive. | **3** |
| 3 | 18215 | but it was a fun run that we had, tobacco and i. kind of like my time in Seattle, it just felt like the fun was over... | **2.8** |
| 4 | 21677 | I have always loved Honda's because they are solid cars, good mileage, great engine, great resale value, and they are dependable. | **2.8** |
| 5 | 24863 | ????????? ? ? ? god bless me the chance to citi final....... | **2.8** |

| | | | |
|---|---|---|---|
| 6 | 33008 | London is a great city, I've always loved visiting, and continue to see it at as amazing place full of amazing people. | **2.8** |
| 7 | 568 | Lol the other day this super ugly fat girl who i have never even talked to before said i looked like a parishiltonwanna-be, i was like wow thanks i love parishilton! | **2.6** |
| 8 | 20659 | Some parts of MIT are really crappy, but some are just amazing, like the Strata Center and the ?Sponge? Went to some info sessions Played tons of videogames and other games while enjoying great food. | **2.6** |
| 9 | 32916 | When the Yankees play the Red Sox, I have to pull for the Bronx Bombers ?that ?s how much I despise Boston ? but then I pray for the Yankee ? s plane to go down. | **2.6** |
| 10 | 3545 | UCLA was awesome = )i had muCHO fun and people were so nice and cool = ) the weather was PERFECTO, lol, but i heard recently it's been as hot as vegas... | **2.4** |

**Table 4: Last 10 Data Set 3 Sentiment Score**

| No | Review ID | Sentiment | Score |
|---|---|---|---|
| 1 | 32858 | The sun is coming up I hate this laptop I hate seattle I hate life I hate my family I hate my friends I hate myself I hate the system I hate the anti-system I hate it all. | **-5.4** |
| 2 | 32148 | and walked away pretending ididnt do it.-visiting bizzle.-driving bitch ass drunk dazy back to chicago at 3 in the morning.-purdue suck shit in football( | **-3.8** |
| 3 | 4126 | i hate i hate i hate i hate i hate i hate san francisco... i totally with appleson, i hate shanghai so much too | **-3.6** |
| 4 | 19759 | hatei hate i hate i hate i hate.. , i hate i ......................................................................................... | **-3.6** |
| 5 | 303 | You are a fucking bitch and I think I may hate you even more than I hate Paris Hilton... | **-3** |
| 6 | 10585 | God I fucking hate Toyota, and their idiot fucking old geezers that work back in the parts department. | **-2.8** |
| 7 | 21645 | Angelina Jolie is a hot bitch for being out of the norm, not afraid ta be real but Jennifer is like, that hot goofy bitch from friends, so it depends on what your in to... | **-2.8** |

| | | | |
|---|---|---|---|
| 8 | 22020 | i really hate parishilton because that bitch bought a grave next to marilynmonroe for her fucking dog... | **-2.8** |
| 9 | 255 | Man if you can't sell angry lefty bullshit in San Francisco, you probably would be challenges to sell ass on a troop ship.... | **-2.4** |
| 10 | 613 | My I hate Forrest Gump and Tom Hanks rant was added to with an I hate Tom Cruise rant Thursday. | **-2.4** |

## DISCUSSIONS

The overall sentiments of MCC Files are generally positive based on the experimental sentiment scoring method used in the study. It is difficult to compare Data Set 1 and Data Set 2 because Google Translate API translated only the words that it can understand. The context of the posts translated using Google Translate API does not carry over from Data Set 1 and Data Set 2. This means that the posts in “TAGLISH” may differ in context when it is translated to the English language using the Google Translate API.

The accuracy of the Word Comparison Algorithm using the declared threshold (90%) was proven efficient in finding the closest word from the posts compared to the AFINN database for the words withfive (5) characters or more.

The sentiment scoring project made for Twitter Posts, from which this project was based, processed faster sentiment scoring compared to the experimental sentiment scoring method because of the following reasons:

- Twitter has a limit of 220 characters per tweet; whereas 60000 characters are allowed in a single post in Facebook;
- Twitter has its own Server Farm to process billions of data at any given time, whereas in this study the researchers used a locally installed web host to process the posts downloaded from MCC Files: And,

- Twitter sentiments scoring using the AFINN database was hard coded in the system; when using the experimental sentiment scoring method in the system, the AFINN words are retrieved every request from the database which can cause overhead delay in the results output.

4783 unique MCC Files posts were processed using data mining technique and the experimental sentiment scoring method in a locally installed web host needed three to four (3-4) hours of processing time to finish the 4783 unique posts.

The experimental sentiment scoring method can be used in different applications such as but not limited to the following:

- Identification of performance evaluation polarity in employee assessments
- Sentiments scoring for essay type activities



The study can be furthered improved by integrating the sentiment scoring analysis to a Facebook API that will automatically score the sentiments of the poster upon submission. Also, the study's language database can be improved by collaborating to language experts that can translate direct to English or direct to Tagalog translation using words or phrases that will enable the context of the post to remain the same even after translation.

## REFERENCES

[1] Xia Hu, Huan Liu, 2012 . Text Analytics in Social Media. In: Mining Text Data pp. 385-414. Springer New York.

[2] K. Shimada, S. Inoue, H. Maeda and T. Endo, 2011. Analyzing Tourism Information on Twitter for a Local City. In the First ACIS International Symposium on Software and Network Engineering, pp: 61–66.

[3] Selvam, B.andAbirami S, 2013. A SURVEY ON OPINION MINING FRAMEWORK. International Journal of Advanced Research in Computer and Communication Engineering., 2(9): 3544-3549

[4] Hejazi H. M., Ahmadi H.andKaramdad F., 2013. The Impact of Information and Computer Technology (ICT) on Staff Organizational Entrepreneurship. J. Basic. Appl. Sci. Res., 3(4): 906-909.

[5] Financial Times, http://www.ft.com/.

[6] Meaningful-Growth, http://blog.twitter.com/2010/12/stocking-stuffer.html.

[7] Why you should care about online reviews? http://loyalrewards.com/images/Online-Review-Infograph-MT.pdf.

[8] CelikyilmazAsli, DilekHakkani-Tur and Junlan Feng, 2010. Probabilistic model-based sentiment analysis of twitter messages. In the Spoken Language Technology Workshop (SLT), pp: 79-84.

[9] Kevin Roebuck, 2011. Sentiment Analysis (ebook).Emereo Pty Limited, pp: 1-2.

[10] Sagolla, D, 2009. 140 Characters: A Style Guide for the Short Form. John Wiley & Sons

[11] Araújo M., Gonçalves P., Cha M., and Benevenuto F, 2014. iFeel: a system that compares and combines sentiment analysis methods. In the Proceedings of the companion publication of the 23rd international conference on World wide web companion, pp: 75-78).

[12] TaboadaMaite et al, 2011. Lexicon-based methods for sentiment analysis. Computational linguistics., 37(2): 267-307.

[13] Turney Peter, 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In the Proceedings of 40th Meeting of the Association for Computational Linguistics, Philadelphia, pp: 417–424.

[14] Pang Bo, Lillian Lee, and ShivakumarVaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques. In the Proceedings of the Conference on Empirical Methods in NLP, Philadelphia, pp: 79–86.

[15] VasfiSisi N. and Derakhshi M. R. F, 2013. Text Classification with Machine Learning Algorithms. J. Basic. Appl. Sci. Res 3(1s)31-35.

[16] Nasukawa Tetsuya and Jeonghee Yi, 2003. Sentiment analysis: Capturing favorability using natural language
processing. In the Proceedings of the KCAP-03, 2nd Intl. Conf. on Knowledge Capture, pp: 70-77.

[17] Dave Kushal, Steve Lawrence and David M. Pennock, 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In the Proceedings of International Conference on World Wide Web, pp: 519-528.

[18] Hatzivassiloglou V. and McKeown K.R., 1997. Predicting the semantic orientation of adjectives. In the Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, pp: 174-181.

[19] Paula Chesley, Bruce Vincent, Li Xu, and Rohini K. Srihari, 2006.Using verbs and adjectives to automatically classify blog sentiment. In the Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to analyzing Weblogs, pp: 27-29.

[20] NamrataGodbole, ManjunathSrinivasaiah, and Steven Skiena, 2007.Large-scale sentiment analysis for news and blogs.In the Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

[21] Adam Bermingham and Alan F. Smeaton, 2010. Classifying sentiment in microblogs: is brevity an advantage? In the Proceedings of the 19th ACM international conference on Information and Knowledge Management, pp: 1833-1836.

[22] Alexander Pak and Patrick Paroubek, 2010. Twitter as a corpus for sentiment analysis and opinion mining. In the Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp: 1320-1326.

[23] Chen Long-Sheng, Cheng-Hsiang Liu, and Hui-Ju Chiu, 2011. A neural network based approach for sentiment classification in the blogosphere. Journal of Informetrics., 5(2): 313-322.

[24] Cohen E.and Krishnamurthy B., 2006. A short walk in the Blogistan. Computer Networks., 50(5): 615–630.

[25] Nardi B. A., Schiano D. J., Gumbrecht M. and Swartz L., 2004. Why we blog? Communications of the ACM, 47(12):41–46

[26] Martin J., 2005. Blogging for Dollars: How Would You Like to Survey 20 Million Consumers in Two Minutes., Fortune, 152(12):S178D.

[27] Microblogging, http://en.wikipedia.org/wiki/Microblogging.

[28] Asghar M. Z.,RahmanUllah, Khan A., Ahmad S.and Nawaz I.U,2014. Political Miner: Opinion Extraction From User Generated Political Reviews. Sci.Int(Lahore),26(1),385-389.

[29] Kumar Akshi, and Teeja Mary Sebastian, 2012. Sentiment analysis: A perspective on its past, present and future. International Journal of Intelligent Systems and Applications (IJISA)., 4(10):1-14.

[30] Manning C. D., RaghavanP.andSchütze H, 2008. Introduction to information retrieval. Cambridge: Cambridge university press, pp: 107-108.

[31] ChaovalitPinvadee and Lina Zhou, 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. In the Proceedings of the 38th Hawaii International Conference on System Sciences, Hawaii, pp: 112c-112c.

[32] Aue Anthony and Michael Gamon, 2005. Customizing sentiment classifiers to new domains: A case study. In the Proceedings of the International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, pp: 1-2.

[33] TaboadaMaite, Julian Brooke and Manfred Stede, 2009.Genre-based paragraph classification for sentiment analysis. In the Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, London, pp:62-70.

[34] Murray Gabriel, ShafiqJoty, Giuseppe Carenini and Raymond Ng, 2008. The University of British Columbia at TAC 2008.Inthe Proceedings of TAC 2008, Gaithersburg, MD.

[35] M. ZubairAsghar, et al. , 2014. Medical Opinion Lexicon: An Incremental Model For Mining HealthReviews. In: International Journal of Academic Research, 6.1.

[36] Tong Richard M, 2001. An operational system for detecting and tracking opinions in on-line discussions. In Working Notes of the ACM SIGIR Workshop on

Operational Text Classification, New York, NY.,pp: 1–6.. Basic. Appl. Sci. Res., 4(6)238-248, 2014

[37] Jinan Fiaidhi et al, 2012. Opinion mining over twitter space: Classifying tweets programmatically using the Rapproach. Digital Information Management (ICDIM), Seventh International Conference on. IEEE, pp: 313-319.

[38] Kumar Akshi, and Teeja Mary Sebastian, 2012.Sentiment Analysis on Twitter. International Journal of Computer Science Issues (IJCSI)., 9(4): 372-378.

[39] K. Manuel, Kishore Varma Indukuri, P. Radha Krishna, 2010. Analyzing Internet Slang for Sentiment Mining, Second Vaagdevi International Conference on Information Technology for Real World Problems., pp: 9-11.

[40] Wang Suge et al, 2011. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Systems with Applications., 38(7): 8696-8702.

[41] Esmin Ahmed Ali Abdalla, R. L. de Oliveira and Stan Matwin, 2012. Hierarchical Classification Approach to Emotion Recognition in Twitter. Machine Learning and Applications (ICMLA)11th International Conference on. IEEE, pp: 381-385.

[42] Bing Liu, Minqing Hu and Junsheng Cheng, 2005. Opinion Observer: Analyzing and Comparing Opinions
on the Web. In the Proceedings of the 14thInternational World Wide Web conference, pp: 342-351.

[43] DadvarMaral and Hauff Claudia and Jong de Franciska, 2011.Scope of negation detection in sentiment analysis. In the Dutch-Belgian Information Retrieval Workshop, pp: 16-19.

[44] Emoticons, http://en.wikipedia.org/wiki/List_of_emoticons

[45] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller, 1990.Introduction to WordNet: An online lexical database. Int. J. Lexicograph., 3(4): 235–244.

[46] Esuli A, Sebastiani F, 2006. Senti Word Net: A Publicly Available Lexical Resource for Opinion Mining. In the Proceedings of International Conference on Language Resources and Evaluation (LREC), Genoa, pp: 417-422.

[47] English-Word List, http://www.curlewcommunications.co.uk/wordlist.html.

[48] Internet Slang Dictionary and Translator, http://www.noslang.com/.

[49] Urban Dictionary, http://www.urbandictionary.com/.

[50] Asghar M. Z., Khan A., Ahmad S.andKundi F. M., 2014. A Review of Feature Extraction in Sentiment Analysis. J. Basic. Appl. Sci. Res., 4(3): 181-186.

[51] M. ZubairAsghar, A. Khan, Sh. Ahmad, F. MasudKundi, 2013. Preprocessing in natural language processing. In: Emerging Issues in the Natural and Applied Sciences pp. 152-161.Progress Publishing Company, Azerbaijan.

[52] Provost Foster J., Tom Fawcett, and Ron Kohavi, 1998.The case against accuracy estimation for comparing induction algorithms. In the Proceedings of the Fifteenth International Conference on Machine Learning, pp: 445-453.

[53] David L. Olson, DursunDelen, 2008. Performance Evaluation for Predictive Modeling. In: Advanced Data Mining Techniques pp: 137-139. Springer-Verlag Berlin Heidelberg.

[54] Matthews Correlation Coefficient, http://en.wikipedia.org/wiki/Matthews_correlation_coefficient

[55] Bin Wem, Wenhua Dai, Junzhe Zhao, 2012. Sentence Sentimental Classification Based on Semantic Comprehension. Computational Intelligence and Design (ISCID), Fifth International Symposium, pp: 458-461.

[56] Chowdhury, Md Faisal Mahbub et al, 2013. Fbk: Sentiment analysis in twitter with tweetsted. Second JointConference on Lexical and Computational Semantics (* SEM): Proceedings of the Seventh International Workshop on Semantic Evaluation, pp: 466-470.

[57] G. Salton and M.J. McGill.Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[58] F.M. Kundi et al. "Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet." Life Science Journal, vol. 11, no. 9, 2014.

[59] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, Michael Etter, "Good Friends, Bad News - Affect and Virality in Twitter", The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011).

# APPENDIX A

## Data Set 1 Sentiment Score

| No | ID | Score | No | ID | Score | No | ID | Score | No | ID | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4173 | 0 | 51 | 446 | -0.6 | 101 | 2270 | 2.6 | 151 | 4411 | 3.2 |
| 2 | 3111 | 4.2 | 52 | 4142 | 1.2 | 102 | 4023 | 1.6 | 152 | 3191 | 0 |
| 3 | 3069 | 1.8 | 53 | 3630 | 1.8 | 103 | 1360 | 1.2 | 153 | 4191 | 0 |
| 4 | 180 | -5.6 | 54 | 4318 | 3.2 | 104 | 4367 | 0 | 154 | 1601 | 2.2 |
| 5 | 3052 | -0.2 | 55 | 4478 | 0.4 | 105 | 764 | 3.4 | 155 | 3082 | 1.8 |
| 6 | 3317 | 8.6 | 56 | 101 | 1.2 | 106 | 682 | 2 | 156 | 3256 | 0.6 |
| 7 | 1531 | 4.6 | 57 | 3084 | 0.8 | 107 | 4271 | 0.4 | 157 | 4448 | 0.8 |
| 8 | 3284 | 1.2 | 58 | 346 | 1.8 | 108 | 4437 | -0.2 | 158 | 2029 | 1 |
| 9 | 3073 | -0.6 | 59 | 2723 | 0.2 | 109 | 4099 | -0.6 | 159 | 4326 | 0 |
| 10 | 4444 | 1.6 | 60 | 4410 | 1.4 | 110 | 873 | 2.8 | 160 | 2297 | 4.2 |
| 11 | 3318 | 3.8 | 61 | 4276 | 3.2 | 111 | 4031 | 0.6 | 161 | 3893 | 0 |
| 12 | 4498 | 0.4 | 62 | 2997 | 0.6 | 112 | 2404 | 0.2 | 162 | 601 | 0.8 |
| 13 | 3026 | 4.2 | 63 | 1058 | 3 | 113 | 1498 | 1.4 | 163 | 3679 | 1.8 |
| 14 | 2996 | 1 | 64 | 3090 | 0.2 | 114 | 4510 | -0.2 | 164 | 1079 | 0.6 |
| 15 | 3211 | 0.8 | 65 | 3314 | 9 | 115 | 4345 | -0.2 | 165 | 3329 | 1.4 |
| 16 | 2171 | 4 | 66 | 3422 | 1 | 116 | 3010 | 0 | 166 | 4316 | 3.6 |
| 17 | 3361 | 10.2 | 67 | 3657 | 0 | 117 | 3070 | 0.2 | 167 | 4317 | 1.2 |
| 18 | 4261 | 19.6 | 68 | 573 | 1.2 | 118 | 3142 | 0 | 168 | 1155 | 8 |
| 19 | 320 | 10.2 | 69 | 1989 | 5.2 | 119 | 283 | 7.6 | 169 | 1340 | -2 |
| 20 | 939 | 0.2 | 70 | 3789 | 0.8 | 120 | 1785 | 2.2 | 170 | 4280 | 0 |
| 21 | 1372 | 0 | 71 | 4483 | 0.6 | 121 | 1355 | -0.2 | 171 | 4300 | 1.8 |
| 22 | 3077 | 2.2 | 72 | 4439 | 2 | 122 | 4465 | 1 | 172 | 3856 | 3 |
| 23 | 4184 | 1.2 | 73 | 374 | 0.6 | 123 | 312 | 4.2 | 173 | 4365 | 0.2 |
| 24 | 3844 | 2.8 | 74 | 4313 | 4.2 | 124 | 3760 | -1 | 174 | 2477 | 3.6 |
| 25 | 1327 | 3.6 | 75 | 3379 | 2.6 | 125 | 2603 | 0 | 175 | 767 | 0.4 |
| 26 | 3792 | 0.2 | 76 | 4084 | 5.8 | 126 | 4426 | 2.8 | 176 | 2391 | 0.2 |
| 27 | 1656 | 0.8 | 77 | 2508 | 1.4 | 127 | 2065 | 0.6 | 177 | 2859 | 0.4 |
| 28 | 130 | 4.4 | 78 | 778 | 1.6 | 128 | 4244 | 2.6 | 178 | 4170 | 3.2 |
| 29 | 3686 | 2.2 | 79 | 505 | 10.8 | 129 | 612 | 1.6 | 179 | 3796 | 3 |
| 30 | 1649 | 3 | 80 | 2640 | 5.8 | 130 | 1302 | 2.2 | 180 | 1425 | -0.6 |
| 31 | 3352 | 2 | 81 | 359 | 0 | 131 | 3078 | 0 | 181 | 1740 | 1.2 |
| 32 | 1725 | 5.2 | 82 | 1623 | 4.8 | 132 | 4234 | 2.8 | 182 | 400 | 0.6 |
| 33 | 3604 | 14.8 | 83 | 3188 | 0.4 | 133 | 2472 | 0 | 183 | 204 | 0 |
| 34 | 1315 | 3.8 | 84 | 1660 | -0.4 | 134 | 96 | 0 | 184 | 4361 | 1.8 |
| 35 | 3051 | 1 | 85 | 4447 | -0.6 | 135 | 4470 | 0.4 | 185 | 386 | 0.4 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 36 | 2615 | **0.6** | 86 | 746 | **0** | 136 | 3396 | **5.4** | 186 | 97 | **2.6** |
| 37 | 593 | **15.8** | 87 | 4427 | **1.8** | 137 | 952 | **-0.6** | 187 | 4155 | **0.6** |
| 38 | 3129 | **1.2** | 88 | 511 | **1.4** | 138 | 1944 | **7.2** | 188 | 2057 | **1.8** |
| 39 | 1599 | **3.2** | 89 | 2103 | **3.6** | 139 | 660 | **0** | 189 | 4053 | **4.2** |
| 40 | 2175 | **3.2** | 90 | 2383 | **-1** | 140 | 674 | **-0.6** | 190 | 1756 | **2.4** |
| 41 | 1197 | **3.6** | 91 | 1299 | **1.6** | 141 | 903 | **1.4** | 191 | 651 | **1.8** |
| 42 | 4093 | **0.6** | 92 | 3083 | **1.6** | 142 | 4223 | **0** | 192 | 236 | **1.6** |
| 43 | 4148 | **0.2** | 93 | 2256 | **0.6** | 143 | 4384 | **1.6** | 193 | 4329 | **1.6** |
| 44 | 3056 | **-1.8** | 94 | 4513 | **3.6** | 144 | 2768 | **-3.8** | 194 | 2767 | **1.8** |
| 45 | 3606 | **1.4** | 95 | 4301 | **1.2** | 145 | 1307 | **1.4** | 195 | 4432 | **2.6** |
| 46 | 4360 | **0** | 96 | 4404 | **5.6** | 146 | 1293 | **2.4** | 196 | 2880 | **0.8** |
| 47 | 4363 | **3.8** | 97 | 503 | **2.4** | 147 | 3014 | **4** | 197 | 4453 | **0.8** |
| 48 | 2104 | **4.8** | 98 | 4473 | **0.8** | 148 | 1426 | **-0.2** | 198 | 335 | **5.8** |
| 49 | 4002 | **3** | 99 | 3734 | **-1.2** | 149 | 3808 | **-0.2** | 199 | 1055 | **5.2** |
| 50 | 3946 | **-0.6** | 100 | 4047 | **1.8** | 150 | 1424 | **-0.4** | 200 | 358 | **0.2** |

# APPENDIX B

## Data Set 2 Sentiment Score

| No | ID | Score | No | ID | Score | No | ID | Score | No | ID | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4173 | 8.2 | 51 | 446 | 1.2 | 101 | 2270 | 0.6 | 151 | 4411 | 8.6 |
| 2 | 3111 | 2.6 | 52 | 4142 | 6.4 | 102 | 4023 | 2 | 152 | 3191 | 10 |
| 3 | 3069 | 14.2 | 53 | 3630 | 4 | 103 | 1360 | -1.4 | 153 | 4191 | 5.4 |
| 4 | 180 | -4.4 | 54 | 4318 | 12.4 | 104 | 4367 | 3.6 | 154 | 1601 | 4.8 |
| 5 | 3052 | -1.2 | 55 | 4478 | 3.6 | 105 | 764 | 5 | 155 | 3082 | 6.2 |
| 6 | 3317 | 9.6 | 56 | 101 | 7.6 | 106 | 682 | 0.8 | 156 | 3256 | 4 |
| 7 | 1531 | 11 | 57 | 3084 | 23.6 | 107 | 4271 | 4.6 | 157 | 4448 | 16 |
| 8 | 3284 | 4.8 | 58 | 346 | 5.8 | 108 | 4437 | 7.2 | 158 | 2029 | 2.2 |
| 9 | 3073 | 5.8 | 59 | 2723 | 0.2 | 109 | 4099 | -7.2 | 159 | 4326 | 9 |
| 10 | 4444 | 4 | 60 | 4410 | 1.4 | 110 | 873 | 6.2 | 160 | 2297 | 7.6 |
| 11 | 3318 | 8 | 61 | 4276 | 3.6 | 111 | 4031 | -1.8 | 161 | 3893 | -0.8 |
| 12 | 4498 | 4.8 | 62 | 2997 | 0.2 | 112 | 2404 | -1.2 | 162 | 601 | 0.6 |
| 13 | 3026 | 20.4 | 63 | 1058 | 3.8 | 113 | 1498 | 5 | 163 | 3679 | 2.6 |
| 14 | 2996 | 5.4 | 64 | 3090 | 2.2 | 114 | 4510 | 15.6 | 164 | 1079 | -3 |
| 15 | 3211 | 1.2 | 65 | 3314 | 7.6 | 115 | 4345 | 1.4 | 165 | 3329 | 1 |
| 16 | 2171 | 7 | 66 | 3422 | 4 | 116 | 3010 | 1 | 166 | 4316 | 9.6 |
| 17 | 3361 | 13.6 | 67 | 3657 | -2 | 117 | 3070 | 4.2 | 167 | 4317 | 15.6 |
| 18 | 4261 | 19.6 | 68 | 573 | 2 | 118 | 3142 | 7 | 168 | 1155 | 8.4 |
| 19 | 320 | 11.8 | 69 | 1989 | 5.2 | 119 | 283 | 6.6 | 169 | 1340 | -0.8 |
| 20 | 939 | 13 | 70 | 3789 | 7.2 | 120 | 1785 | 7.4 | 170 | 4280 | 1.2 |
| 21 | 1372 | 3.6 | 71 | 4483 | -10.4 | 121 | 1355 | 4.6 | 171 | 4300 | 9.8 |
| 22 | 3077 | -2.2 | 72 | 4439 | 10.8 | 122 | 4465 | 9.8 | 172 | 3856 | 1.2 |
| 23 | 4184 | 4.4 | 73 | 374 | 1.2 | 123 | 312 | 4.6 | 173 | 4365 | -2.8 |
| 24 | 3844 | 4.6 | 74 | 4313 | 17.4 | 124 | 3760 | 0.8 | 174 | 2477 | 15 |
| 25 | 1327 | 3.2 | 75 | 3379 | -3.2 | 125 | 2603 | 9.4 | 175 | 767 | 3.6 |
| 26 | 3792 | 1.2 | 76 | 4084 | 12.6 | 126 | 4426 | 6.6 | 176 | 2391 | 0.8 |
| 27 | 1656 | 6.2 | 77 | 2508 | 3.2 | 127 | 2065 | -3.8 | 177 | 2859 | -5.4 |
| 28 | 130 | 9 | 78 | 778 | 6.6 | 128 | 4244 | 6.6 | 178 | 4170 | 9.2 |
| 29 | 3686 | 4.8 | 79 | 505 | 11 | 129 | 612 | 5.2 | 179 | 3796 | 22.8 |
| 30 | 1649 | 5.6 | 80 | 2640 | 6.4 | 130 | 1302 | 9.8 | 180 | 1425 | -0.2 |
| 31 | 3352 | 3.6 | 81 | 359 | -2.4 | 131 | 3078 | 0 | 181 | 1740 | 9.6 |
| 32 | 1725 | 3.2 | 82 | 1623 | 11 | 132 | 4234 | 6 | 182 | 400 | -3.4 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | 3604 | 13.4 | 83 | 3188 | -0.2 | 133 | 2472 | 1 | 183 | 204 | 14.6 |
| 34 | 1315 | 7.4 | 84 | 1660 | 5 | 134 | 96 | 9.6 | 184 | 4361 | 0.6 |
| 35 | 3051 | -2 | 85 | 4447 | 3 | 135 | 4470 | 5.6 | 185 | 386 | -4 |
| 36 | 2615 | 6.4 | 86 | 746 | -1 | 136 | 3396 | 5.8 | 186 | 97 | 7.4 |
| 37 | 593 | 17.4 | 87 | 4427 | 1.2 | 137 | 952 | 4.8 | 187 | 4155 | 6 |
| 38 | 3129 | 5.4 | 88 | 511 | 3.2 | 138 | 1944 | 10 | 188 | 2057 | 0.8 |
| 39 | 1599 | -2.4 | 89 | 2103 | 5.2 | 139 | 660 | 1.4 | 189 | 4053 | 10.2 |
| 40 | 2175 | 9.6 | 90 | 2383 | 0.6 | 140 | 674 | 0 | 190 | 1756 | 6.4 |
| 41 | 1197 | 10.6 | 91 | 1299 | 4.4 | 141 | 903 | 5.8 | 191 | 651 | -0.2 |
| 42 | 4093 | 3.8 | 92 | 3083 | 8.4 | 142 | 4223 | 18.2 | 192 | 236 | 3 |
| 43 | 4148 | 1.8 | 93 | 2256 | 6.2 | 143 | 4384 | 0.4 | 193 | 4329 | 1.4 |
| 44 | 3056 | -4.8 | 94 | 4513 | 4.8 | 144 | 2768 | -10.6 | 194 | 2767 | 1.4 |
| 45 | 3606 | 5 | 95 | 4301 | 4 | 145 | 1307 | 22.4 | 195 | 4432 | 3 |
| 46 | 4360 | 3.4 | 96 | 4404 | 5.6 | 146 | 1293 | 6.8 | 196 | 2880 | -0.2 |
| 47 | 4363 | 2.4 | 97 | 503 | 2.6 | 147 | 3014 | 3.2 | 197 | 4453 | 0 |
| 48 | 2104 | 4.4 | 98 | 4473 | 4.8 | 148 | 1426 | 2.2 | 198 | 335 | 5 |
| 49 | 4002 | 2.8 | 99 | 3734 | 0 | 149 | 3808 | -2.8 | 199 | 1055 | 7.4 |
| 50 | 3946 | 5.2 | 100 | 4047 | -2.6 | 150 | 1424 | -2 | 200 | 358 | -1.2 |

## APPENDIC C

### Comparison Scores of Data Set 1 and Data Set 2

| No | ID Data Set1 | Data Set 1 | ID Data Set2 | Data Set 2 | Difference (Translated - Original) |
|---|---|---|---|---|---|
| 1 | 4173 | 0 | 4173 | 8.2 | **8.2** |
| 2 | 3111 | 4.2 | 3111 | 2.6 | **-1.6** |
| 3 | 3069 | 1.8 | 3069 | 14.2 | **12.4** |
| 4 | 180 | -5.6 | 180 | -4.4 | **1.2** |
| 5 | 3052 | -0.2 | 3052 | -1.2 | **-1** |
| 6 | 3317 | 8.6 | 3317 | 9.6 | **1** |
| 7 | 1531 | 4.6 | 1531 | 11 | **6.4** |
| 8 | 3284 | 1.2 | 3284 | 4.8 | **3.6** |
| 9 | 3073 | -0.6 | 3073 | 5.8 | **6.4** |
| 10 | 4444 | 1.6 | 4444 | 4 | **2.4** |
| 11 | 3318 | 3.8 | 3318 | 8 | **4.2** |
| 12 | 4498 | 0.4 | 4498 | 4.8 | **4.4** |
| 13 | 3026 | 4.2 | 3026 | 20.4 | **16.2** |
| 14 | 2996 | 1 | 2996 | 5.4 | **4.4** |
| 15 | 3211 | 0.8 | 3211 | 1.2 | **0.4** |
| 16 | 2171 | 4 | 2171 | 7 | **3** |
| 17 | 3361 | 10.2 | 3361 | 13.6 | **3.4** |
| 18 | 4261 | 19.6 | 4261 | 19.6 | **0** |
| 19 | 320 | 10.2 | 320 | 11.8 | **1.6** |
| 20 | 939 | 0.2 | 939 | 13 | **12.8** |
| 21 | 1372 | 0 | 1372 | 3.6 | **3.6** |
| 22 | 3077 | 2.2 | 3077 | -2.2 | **-4.4** |
| 23 | 4184 | 1.2 | 4184 | 4.4 | **3.2** |
| 24 | 3844 | 2.8 | 3844 | 4.6 | **1.8** |
| 25 | 1327 | 3.6 | 1327 | 3.2 | **-0.4** |
| 26 | 3792 | 0.2 | 3792 | 1.2 | **1** |
| 27 | 1656 | 0.8 | 1656 | 6.2 | **5.4** |
| 28 | 130 | 4.4 | 130 | 9 | **4.6** |
| 29 | 3686 | 2.2 | 3686 | 4.8 | **2.6** |
| 30 | 1649 | 3 | 1649 | 5.6 | **2.6** |
| 31 | 3352 | 2 | 3352 | 3.6 | **1.6** |
| 32 | 1725 | 5.2 | 1725 | 3.2 | **-2** |

| | | | | | |
|---|---|---|---|---|---|
| 33 | 3604 | 14.8 | 3604 | 13.4 | **-1.4** |
| 34 | 1315 | 3.8 | 1315 | 7.4 | **3.6** |
| 35 | 3051 | 1 | 3051 | -2 | **-3** |
| 36 | 2615 | 0.6 | 2615 | 6.4 | **5.8** |
| 37 | 593 | 15.8 | 593 | 17.4 | **1.6** |
| 38 | 3129 | 1.2 | 3129 | 5.4 | **4.2** |
| 39 | 1599 | 3.2 | 1599 | -2.4 | **-5.6** |
| 40 | 2175 | 3.2 | 2175 | 9.6 | **6.4** |
| 41 | 1197 | 3.6 | 1197 | 10.6 | **7** |
| 42 | 4093 | 0.6 | 4093 | 3.8 | **3.2** |
| 43 | 4148 | 0.2 | 4148 | 1.8 | **1.6** |
| 44 | 3056 | -1.8 | 3056 | -4.8 | **-3** |
| 45 | 3606 | 1.4 | 3606 | 5 | **3.6** |
| 46 | 4360 | 0 | 4360 | 3.4 | **3.4** |
| 47 | 4363 | 3.8 | 4363 | 2.4 | **-1.4** |
| 48 | 2104 | 4.8 | 2104 | 4.4 | **-0.4** |
| 49 | 4002 | 3 | 4002 | 2.8 | **-0.2** |
| 50 | 3946 | -0.6 | 3946 | 5.2 | **5.8** |
| 51 | 446 | -0.6 | 446 | 1.2 | **1.8** |
| 52 | 4142 | 1.2 | 4142 | 6.4 | **5.2** |
| 53 | 3630 | 1.8 | 3630 | 4 | **2.2** |
| 54 | 4318 | 3.2 | 4318 | 12.4 | **9.2** |
| 55 | 4478 | 0.4 | 4478 | 3.6 | **3.2** |
| 56 | 101 | 1.2 | 101 | 7.6 | **6.4** |
| 57 | 3084 | 0.8 | 3084 | 23.6 | **22.8** |
| 58 | 346 | 1.8 | 346 | 5.8 | **4** |
| 59 | 2723 | 0.2 | 2723 | 0.2 | **0** |
| 60 | 4410 | 1.4 | 4410 | 1.4 | **0** |
| 61 | 4276 | 3.2 | 4276 | 3.6 | **0.4** |
| 62 | 2997 | 0.6 | 2997 | 0.2 | **-0.4** |
| 63 | 1058 | 3 | 1058 | 3.8 | **0.8** |
| 64 | 3090 | 0.2 | 3090 | 2.2 | **2** |
| 65 | 3314 | 9 | 3314 | 7.6 | **-1.4** |
| 66 | 3422 | 1 | 3422 | 4 | **3** |
| 67 | 3657 | 0 | 3657 | -2 | **-2** |
| 68 | 573 | 1.2 | 573 | 2 | **0.8** |
| 69 | 1989 | 5.2 | 1989 | 5.2 | **0** |
| 70 | 3789 | 0.8 | 3789 | 7.2 | **6.4** |
| 71 | 4483 | 0.6 | 4483 | -10.4 | **-11** |

| | | | | | |
|---|---|---|---|---|---|
| 72 | 4439 | 2 | 4439 | 10.8 | **8.8** |
| 73 | 374 | 0.6 | 374 | 1.2 | **0.6** |
| 74 | 4313 | 4.2 | 4313 | 17.4 | **13.2** |
| 75 | 3379 | 2.6 | 3379 | -3.2 | **-5.8** |
| 76 | 4084 | 5.8 | 4084 | 12.6 | **6.8** |
| 77 | 2508 | 1.4 | 2508 | 3.2 | **1.8** |
| 78 | 778 | 1.6 | 778 | 6.6 | **5** |
| 79 | 505 | 10.8 | 505 | 11 | **0.2** |
| 80 | 2640 | 5.8 | 2640 | 6.4 | **0.6** |
| 81 | 359 | 0 | 359 | -2.4 | **-2.4** |
| 82 | 1623 | 4.8 | 1623 | 11 | **6.2** |
| 83 | 3188 | 0.4 | 3188 | -0.2 | **-0.6** |
| 84 | 1660 | -0.4 | 1660 | 5 | **5.4** |
| 85 | 4447 | -0.6 | 4447 | 3 | **3.6** |
| 86 | 746 | 0 | 746 | -1 | **-1** |
| 87 | 4427 | 1.8 | 4427 | 1.2 | **-0.6** |
| 88 | 511 | 1.4 | 511 | 3.2 | **1.8** |
| 89 | 2103 | 3.6 | 2103 | 5.2 | **1.6** |
| 90 | 2383 | -1 | 2383 | 0.6 | **1.6** |
| 91 | 1299 | 1.6 | 1299 | 4.4 | **2.8** |
| 92 | 3083 | 1.6 | 3083 | 8.4 | **6.8** |
| 93 | 2256 | 0.6 | 2256 | 6.2 | **5.6** |
| 94 | 4513 | 3.6 | 4513 | 4.8 | **1.2** |
| 95 | 4301 | 1.2 | 4301 | 4 | **2.8** |
| 96 | 4404 | 5.6 | 4404 | 5.6 | **0** |
| 97 | 503 | 2.4 | 503 | 2.6 | **0.2** |
| 98 | 4473 | 0.8 | 4473 | 4.8 | **4** |
| 99 | 3734 | -1.2 | 3734 | 0 | **1.2** |
| 100 | 4047 | 1.8 | 4047 | -2.6 | **-4.4** |
| 101 | 2270 | 2.6 | 2270 | 0.6 | **-2** |
| 102 | 4023 | 1.6 | 4023 | 2 | **0.4** |
| 103 | 1360 | 1.2 | 1360 | -1.4 | **-2.6** |
| 104 | 4367 | 0 | 4367 | 3.6 | **3.6** |
| 105 | 764 | 3.4 | 764 | 5 | **1.6** |
| 106 | 682 | 2 | 682 | 0.8 | **-1.2** |
| 107 | 4271 | 0.4 | 4271 | 4.6 | **4.2** |
| 108 | 4437 | -0.2 | 4437 | 7.2 | **7.4** |
| 109 | 4099 | -0.6 | 4099 | -7.2 | **-6.6** |
| 110 | 873 | 2.8 | 873 | 6.2 | **3.4** |

| | | | | | |
|---|---|---|---|---|---|
| 111 | 4031 | 0.6 | 4031 | -1.8 | **-2.4** |
| 112 | 2404 | 0.2 | 2404 | -1.2 | **-1.4** |
| 113 | 1498 | 1.4 | 1498 | 5 | **3.6** |
| 114 | 4510 | -0.2 | 4510 | 15.6 | **15.8** |
| 115 | 4345 | -0.2 | 4345 | 1.4 | **1.6** |
| 116 | 3010 | 0 | 3010 | 1 | **1** |
| 117 | 3070 | 0.2 | 3070 | 4.2 | **4** |
| 118 | 3142 | 0 | 3142 | 7 | **7** |
| 119 | 283 | 7.6 | 283 | 6.6 | **-1** |
| 120 | 1785 | 2.2 | 1785 | 7.4 | **5.2** |
| 121 | 1355 | -0.2 | 1355 | 4.6 | **4.8** |
| 122 | 4465 | 1 | 4465 | 9.8 | **8.8** |
| 123 | 312 | 4.2 | 312 | 4.6 | **0.4** |
| 124 | 3760 | -1 | 3760 | 0.8 | **1.8** |
| 125 | 2603 | 0 | 2603 | 9.4 | **9.4** |
| 126 | 4426 | 2.8 | 4426 | 6.6 | **3.8** |
| 127 | 2065 | 0.6 | 2065 | -3.8 | **-4.4** |
| 128 | 4244 | 2.6 | 4244 | 6.6 | **4** |
| 129 | 612 | 1.6 | 612 | 5.2 | **3.6** |
| 130 | 1302 | 2.2 | 1302 | 9.8 | **7.6** |
| 131 | 3078 | 0 | 3078 | 0 | **0** |
| 132 | 4234 | 2.8 | 4234 | 6 | **3.2** |
| 133 | 2472 | 0 | 2472 | 1 | **1** |
| 134 | 96 | 0 | 96 | 9.6 | **9.6** |
| 135 | 4470 | 0.4 | 4470 | 5.6 | **5.2** |
| 136 | 3396 | 5.4 | 3396 | 5.8 | **0.4** |
| 137 | 952 | -0.6 | 952 | 4.8 | **5.4** |
| 138 | 1944 | 7.2 | 1944 | 10 | **2.8** |
| 139 | 660 | 0 | 660 | 1.4 | **1.4** |
| 140 | 674 | -0.6 | 674 | 0 | **0.6** |
| 141 | 903 | 1.4 | 903 | 5.8 | **4.4** |
| 142 | 4223 | 0 | 4223 | 18.2 | **18.2** |
| 143 | 4384 | 1.6 | 4384 | 0.4 | **-1.2** |
| 144 | 2768 | -3.8 | 2768 | -10.6 | **-6.8** |
| 145 | 1307 | 1.4 | 1307 | 22.4 | **21** |
| 146 | 1293 | 2.4 | 1293 | 6.8 | **4.4** |
| 147 | 3014 | 4 | 3014 | 3.2 | **-0.8** |
| 148 | 1426 | -0.2 | 1426 | 2.2 | **2.4** |
| 149 | 3808 | -0.2 | 3808 | -2.8 | **-2.6** |

| | | | | | |
|---|---|---|---|---|---|
| 150 | 1424 | -0.4 | 1424 | -2 | **-1.6** |
| 151 | 4411 | 3.2 | 4411 | 8.6 | **5.4** |
| 152 | 3191 | 0 | 3191 | 10 | **10** |
| 153 | 4191 | 0 | 4191 | 5.4 | **5.4** |
| 154 | 1601 | 2.2 | 1601 | 4.8 | **2.6** |
| 155 | 3082 | 1.8 | 3082 | 6.2 | **4.4** |
| 156 | 3256 | 0.6 | 3256 | 4 | **3.4** |
| 157 | 4448 | 0.8 | 4448 | 16 | **15.2** |
| 158 | 2029 | 1 | 2029 | 2.2 | **1.2** |
| 159 | 4326 | 0 | 4326 | 9 | **9** |
| 160 | 2297 | 4.2 | 2297 | 7.6 | **3.4** |
| 161 | 3893 | 0 | 3893 | -0.8 | **-0.8** |
| 162 | 601 | 0.8 | 601 | 0.6 | **-0.2** |
| 163 | 3679 | 1.8 | 3679 | 2.6 | **0.8** |
| 164 | 1079 | 0.6 | 1079 | -3 | **-3.6** |
| 165 | 3329 | 1.4 | 3329 | 1 | **-0.4** |
| 166 | 4316 | 3.6 | 4316 | 9.6 | **6** |
| 167 | 4317 | 1.2 | 4317 | 15.6 | **14.4** |
| 168 | 1155 | 8 | 1155 | 8.4 | **0.4** |
| 169 | 1340 | -2 | 1340 | -0.8 | **1.2** |
| 170 | 4280 | 0 | 4280 | 1.2 | **1.2** |
| 171 | 4300 | 1.8 | 4300 | 9.8 | **8** |
| 172 | 3856 | 3 | 3856 | 1.2 | **-1.8** |
| 173 | 4365 | 0.2 | 4365 | -2.8 | **-3** |
| 174 | 2477 | 3.6 | 2477 | 15 | **11.4** |
| 175 | 767 | 0.4 | 767 | 3.6 | **3.2** |
| 176 | 2391 | 0.2 | 2391 | 0.8 | **0.6** |
| 177 | 2859 | 0.4 | 2859 | -5.4 | **-5.8** |
| 178 | 4170 | 3.2 | 4170 | 9.2 | **6** |
| 179 | 3796 | 3 | 3796 | 22.8 | **19.8** |
| 180 | 1425 | -0.6 | 1425 | -0.2 | **0.4** |
| 181 | 1740 | 1.2 | 1740 | 9.6 | **8.4** |
| 182 | 400 | 0.6 | 400 | -3.4 | **-4** |
| 183 | 204 | 0 | 204 | 14.6 | **14.6** |
| 184 | 4361 | 1.8 | 4361 | 0.6 | **-1.2** |
| 185 | 386 | 0.4 | 386 | -4 | **-4.4** |
| 186 | 97 | 2.6 | 97 | 7.4 | **4.8** |
| 187 | 4155 | 0.6 | 4155 | 6 | **5.4** |
| 188 | 2057 | 1.8 | 2057 | 0.8 | **-1** |

| | | | | | |
|---|---|---|---|---|---|
| 189 | 4053 | 4.2 | 4053 | 10.2 | **6** |
| 190 | 1756 | 2.4 | 1756 | 6.4 | **4** |
| 191 | 651 | 1.8 | 651 | -0.2 | **-2** |
| 192 | 236 | 1.6 | 236 | 3 | **1.4** |
| 193 | 4329 | 1.6 | 4329 | 1.4 | **-0.2** |
| 194 | 2767 | 1.8 | 2767 | 1.4 | **-0.4** |
| 195 | 4432 | 2.6 | 4432 | 3 | **0.4** |
| 196 | 2880 | 0.8 | 2880 | -0.2 | **-1** |
| 197 | 4453 | 0.8 | 4453 | 0 | **-0.8** |
| 198 | 335 | 5.8 | 335 | 5 | **-0.8** |
| 199 | 1055 | 5.2 | 1055 | 7.4 | **2.2** |
| 200 | 358 | 0.2 | 358 | -1.2 | **-1.4** |