Using Clustering and Sentiment Analysis on Twitter

GRADUATE PROJECT REPORT

Submitted to the Faculty of
The School of Engineering & Computing Sciences
Texas A&M University-Corpus Christi
Corpus Christi, TX

in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Science

by

Ming-Hsuan Wu
Fall 2014

Committee Members

Dr. Longzhuang Li ______________________________
Committee Chairperson

Dr. David Thomas ______________________________
Committee Member

ABSTRACT

Recently, social media has become important for social networking and content sharing. Twitter, an online social network, allows users to upload short text messages, also known as tweets, with up to 140 characters. A lot of people use sentiment analysis on Twitter to do opinion mining. People choose Twitter because Twitter serves as a good platform for sentiment analysis because of its large user base from different sociocultural zones. The objective of Sentiment Analysis is to identify any clue of positive or negative emotions in a piece of text reflective of the authors' opinions on a subject.

Twitter API, twitter4j, is processed to search selected popular electronic products on Twitter. K-means cluster approach is used to find some clusters that have similar sentences. Similar sentence means the sentences have the same keywords. It means the tweets in the cluster are about how people think about similar features of selected popular electronic products. Each cluster is entered into feature-based sentiment analysis to get the score. After that, the total tweets also process in the sentiment analysis system to analyze how people think about selected popular electronic products. The system uses TF-IDF, k-means algorithm, SentiWordNet and Stanford tool to handle different level steps.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Twitter is a microblogging website that has become increasingly popular with the network community. Users update short messages, also known as Tweets, which are limited to 140 characters. Users frequently share their personal opinions on many subjects, discuss current topics and write about life events. This platform is favored by many users because it is free from political and economic limitations and is easily available to millions of people. As the amount of users increase, microblogging platforms are becoming a place to find strong viewpoints and sentiment.

People use twitter to predict a lot of different areas. For example, people have already predicted the stock market success by using data from Twitter [1]. People use Twitter to forecast box-office revenues for movies [2]. From these case studies, we can know that Twitter is really useful for predicting products, services, or markets. It is one important reason why Twitter is chosen to predict how people think about the popularity of electronic products. Another reason is because Twitter serves as a worthy platform for sentiment analysis due to its large user base from a variety of social and cultural regions worldwide. Twitter contains a vast number of tweets, with millions being added every day. This can be easily collected through its APIs (Application Program Interface), which makes it easy to build a great training set.

2. BACKGROUND AND RATIONALE

2.1 Sentiment Computing and Classification

Sina Weibo is a Chinese microblogging website, similar to Twitter, which allows users to post with a 140-character limit, mention or talk to other people using "@UserName" format, add hashtags with "#HashName#" format. The Weibo is one of the most popular sites in China, in use by well over 30% of Internet users, with a market penetration similar to the United States' Twitter [3].

This approach builds a Sentiment Dictionary by using the Word2vec tool, which is modeled after the Semantic Orientation Pointwise Similarity Distance (SO-SD) model [4]. Once this step is completed, the Emotional Dictionary is used to get the emotional trends from messages posted by users on Weibo. In this approach, Weibo contents are categorized into three groups: positive, negative and neutral. After the grouping has been completed, the approach uses the Paoding word-segmentation tool to separate Weibo contents into different Chinese words. Next, 70% of the processed words from Weibo are used to train the Word2vec tool and this gets an extended Weibo Sentiment Dictionary. The remaining 30% of words are used to confirm the success of the approach. Last, Weibo Sentiment Dictionary is used to estimate the Weibo sentiment trends. Figure 2.1 illustrates the steps in this approach.

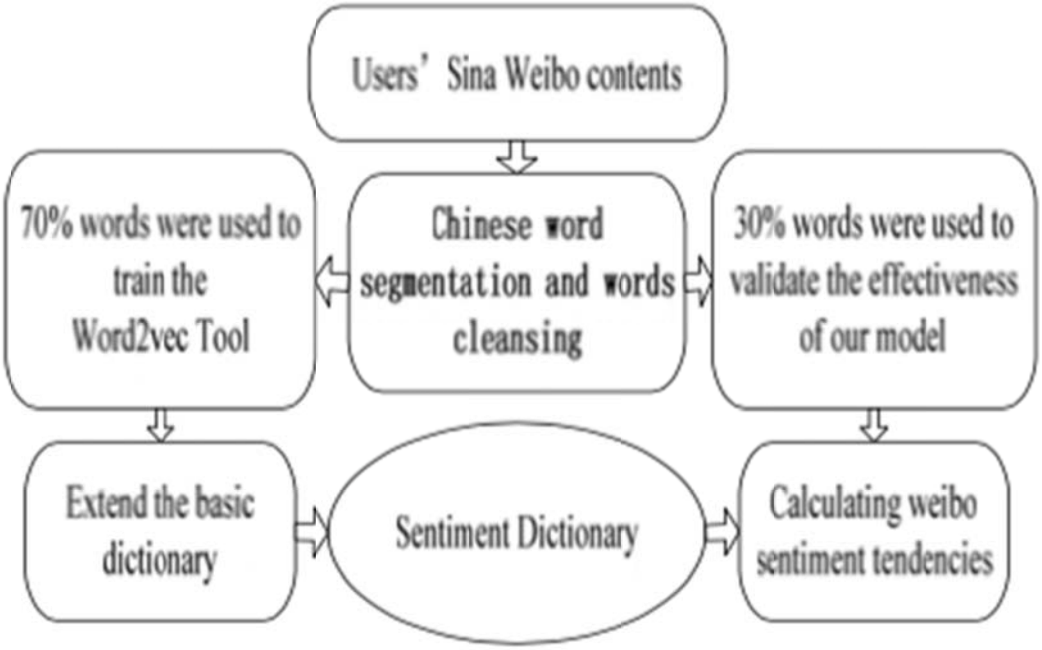

Users' Sina Weibo contents
70% words were used to train the Word2vec Tool
Chinese word segmentation and words cleansing
30% words were used to validate the effectiveness of our model
Extend the basic dictionary
Sentiment Dictionary
Calculating weibo sentiment tendencies

Figure 2.1. Sentiment Computing and Classification [3]

An easy way to examine the resulting depictions from this is to find a closely related word or common synonym for the word specified by the user. The distance tool helps to complete this task. For example, if you enter 'Boston', the distance tool displays the most closely related words and their distances to 'Boston'.

This approach allows for 70% of the collected words to be used to train the Word2vec tool. The remaining 30% of collected words are used to estimate the Weibo sentiment trends. The most useful data is not enough because there is so much data that is used to extend the basic dictionary.

2.2 Clustering

One of the issues with Twitter is that users post many opinions and these opinions are broad. Users discuss many different topics in their posts, so these posts focus on more than just the product review. Based on this knowledge, the collection of such wildranging data would result in inaccurate data, which is reason clustering is necessary to use first in order to help discover data with similarities. Clustering can be considered the most important machine learning problem. It is the task of grouping a set of objects in such a way that objects in the same group are called a cluster [5]. The clusters are more similar to each other than to those in other clusters.

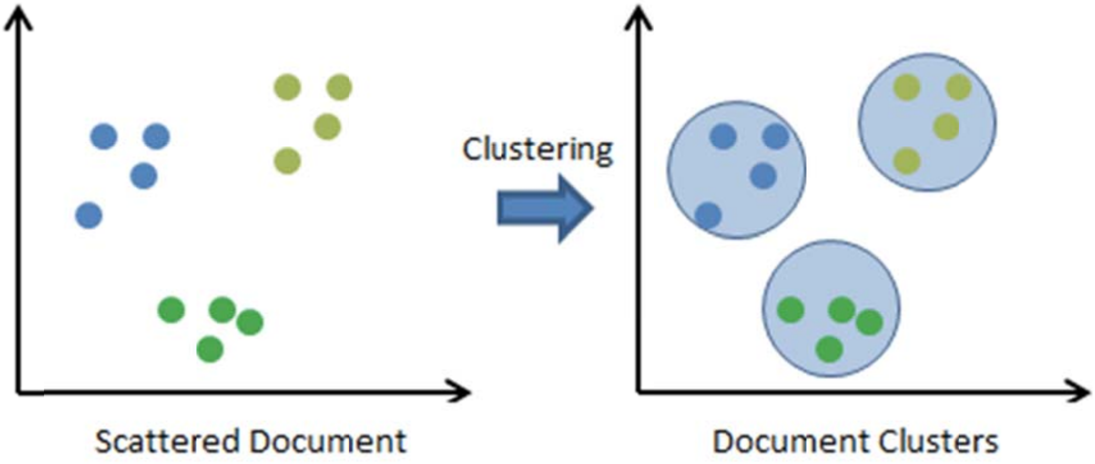


Figure 2.2. Clustering [6]

In Figure 2.2, we can easily separate data to 3 clusters. Distance is an important point to know because each object should belong to a cluster. Two or more objects belong to the same cluster if they are close, according to the distance.

2.2.1 Twitter Clusters System

Figure 2.3 shows the whole design for the method. In order to apply this method, there is a set of steps to be followed. First, eight Twitter feeds must be selected so that all tweets are in English and probable to create clusters. Second, 9 days out of a two months time frame, approximately 1000 tweets is collected. Third, the Tweets must be organized and the tweets with a minimum of 60 characters that are similar are removed in order to prevent repetition in news tweeted.

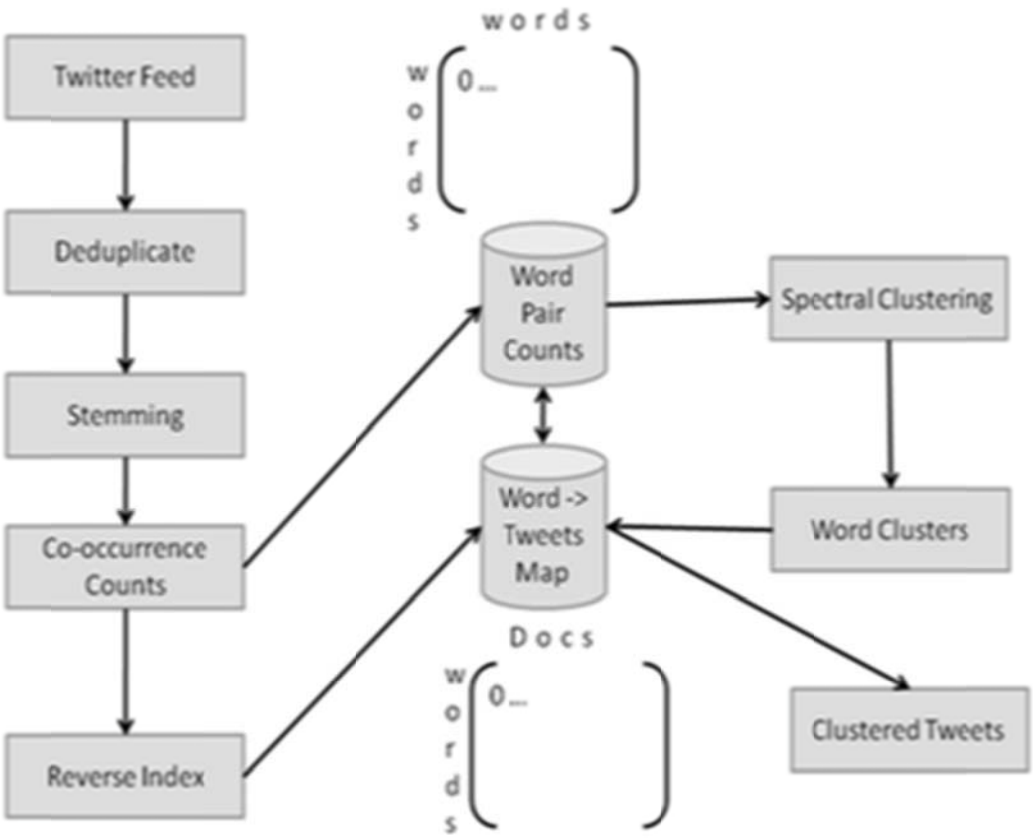


Figure 2.3. Twitter Clusters System Design [7]

Fourth, spaces must be added around punctuation such as , ; : - but not . ’ because splitting words such as “U.S.” or “don’t” is not wanted. Fifth, basic stop words and specific twitter stop words such as “alert” and “breaking” need to be removed. Sixth, to help in clustering, if we care about the word clusters making sense and maybe use them for search, we should avoid stemming. Seventh, with these features, a “word co-occurrence matrix W” can be created. W_{ij} is set to *n*, if there are *n* tweets that contain both the features *i* and *j*. After that, the weight matrix needs to be used to perform “spectral clustering” using W to get word clusters. Last, in addition to using the word, use the reverse index to get tweet clusters. [7].

Unfortunately, the negative side to using this method is that too much time is taken for data collection. Furthermore, this method using clustering too much, which adds more to the amount of time used. Most of the time, clustering time consumption is focused on finding a good center point. Therefore, less clustering is usually a better choice when trying to save time.

2.2.2 K-means Algorithm

The k-means clustering algorithm is known to be efficient in clustering large data sets, and is one of the simplest and the best known machine learning algorithms that solve the well-known clustering problem [8].

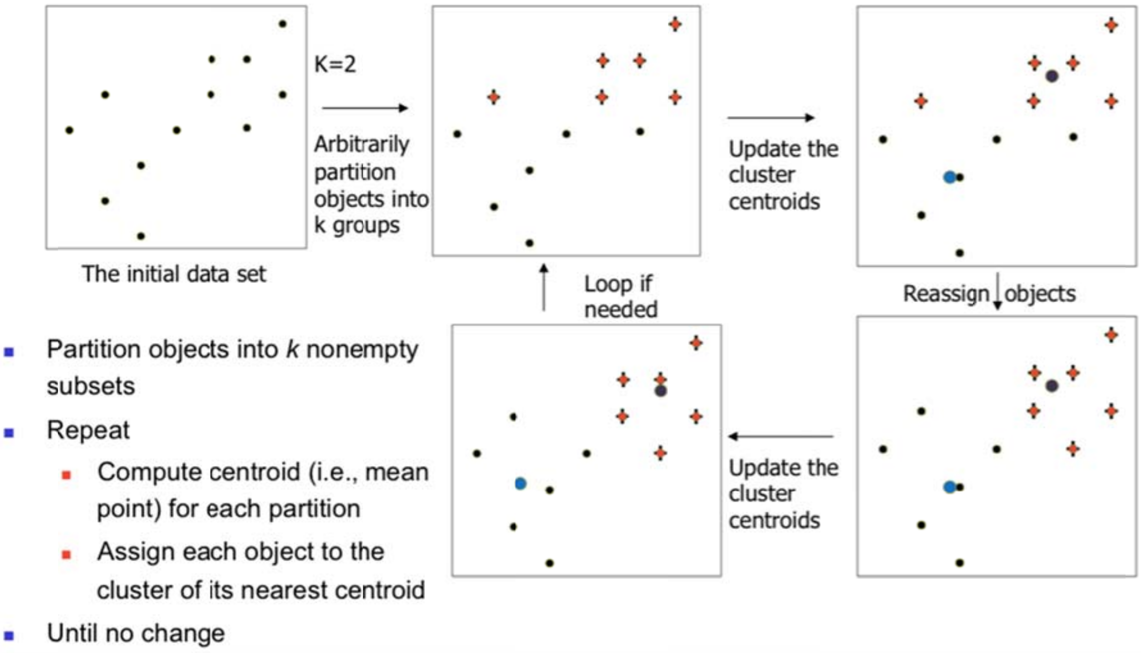

K=2
Arbitrarily partition objects into k groups
The initial data set
Update the cluster centroids
Reassign objects
Update the cluster centroids
Loop if needed
Partition objects into *k* nonempty subsets
Repeat
Compute centroid (i.e., mean point) for each partition
Assign each object to the cluster of its nearest centroid
Until no change

Figure 2.4. K-means Algorithm [9]

The Figure 2.4 shows the four steps of the k-means algorithm. The first step divides items into *k* nonempty subgroups. In the second, the compute seed points to the centroids of the clusters of the current divisions. The centroid is at the center, which means the middle point of the cluster group. The third step is when each object is assigned to the cluster with the nearest seed point. The fourth and last step goes back to Step 2 and stops when the assignment does not change [9].

The positive side for k-means is the simplest. All you need to do is choose k and run it a number of times, especially if the clusters are circular shape. Most of people do not need a complex cluster algorithm.

K-means process has some weaknesses. First, there is a problem with comparing the quality of the clusters. Second, because there is a fixed number of cluster, it can be hard to find out what K should be. Third, k-means only work well with circular cluster shape. Fourth, when the original partitions are not the same, this may cause final clusters that are also different. It is useful to run the program again by like and unlike K values, to compare the outcomes gained [9].

2.3 Sentiment Analysis

Sentiment analysis, also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards things such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they are now all under the authority of sentiment analysis or opinion mining [10].

We can know how users feel about a product or service and this can help, especially in business decisions for corporates with sentiment analysis. Also, political parties and social organizations can collect feedback about their programs. Furthermore, entertainers such as actors, musicians, and artists can connect with their fans and find the viewpoints on their work. Mostly, this can act as an automatic surveying method, which does not require manual entry [11].

2.4 Feature-based Sentiment Analysis

The document of people's opinions is from the paragraphs, the paragraph is from the sentences, the sentence is from the words. Therefore, the first feature that feature-based sentiment analysis models discover is the word in a sentence. It determines if the opinions are positive, negative or neutral. The opinions can be about a topic, event, product, service, etc. Sentiment analysis separates document into paragraphs and then separate paragraph into sentences. After that, sentences are separated into words. In the next step, sentiment analysis forces feature from word-level, sentence-level, paragraph-level, to document-level. Once this is complete, calculate the positive score, negative score, or neutral score from each level and add the final score together. Finally, change the opinion to number, and analyze the number to understand how people's real thinking is.

This feature-based sentiment analysis system uses Stanford tool and SentiWordNet [12]. SentiWordNet is a resource for supporting opinion mining applications. SentiWordNet relates to the positive, negative, and neutral opinions to tag all the WordNet synsets [13]. It has two steps: preparing data and building processing components [14]. First, this system uses SentiWordNet to create positive and negative words lists, and lists with words that can reverse, increase or decrease the opinion.

Second, this system uses the processing components and enters text files from Twitter to find the product and the comments. This system uses an open source tool called Stanford for stemming and tagging the parts-of-speech.

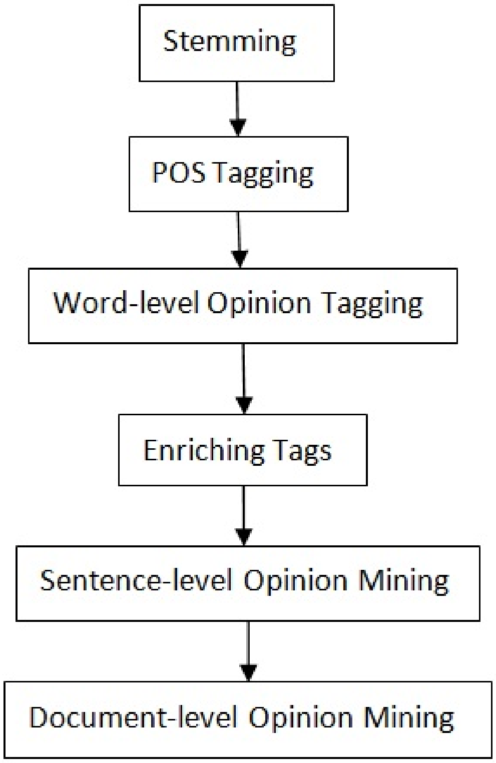

Stemming
POS Tagging
Word-level Opinion Tagging
Enriching Tags
Sentence-level Opinion Mining
Document-level Opinion Mining

Figure 2.5. Flow Diagram of the Proposed System [14]

First, the Stemming part is when all data from the text document is collected. Second, the Stanford POS Tagger is used to do the POS Tagging [15]. Third, the SentiWordNet 3.0 is used to make the positive and negative word lists. Fourth, the Enriching tag is used as the special tags for reversed word lists. For example, negation Neg is positive. The increase and decrease words are tagged to increase the opinion and/or decrease the opinion. Fifth, sentence-level opinion mining sets all opinion values to begin at 0. The lpos, pos, vpo are +1, +2, +3. The lneg, neg, vneg are -1, -2, -3. For example, good and easy to use are +2. Bad and hard to use are -2. Next, calculate the score by using sentence-level opinion combination methods. Last, add all totals of sentence-level opinion together. There has a table to verify if the opinion text is positive or negative. For instance, if the final score is more than 60%, this shows a strong positive. However, if the final total is less than -60%, this shows a strong negative. For example, I want to analyze a sentence: this phone is good and easy to use, and the sentence becomes after process:

This/[POS_DT|Stm_this] phone/[POS_NN|Stm_phone] is/[POS_VBZ|Stm_be] good/[POS_JJ|Stm_good|Opn_positive|pos] and/ [POS_CC|Stm_and] easy/[POS_JJ|Stm_easy|pos] to/[POS_TO|Stm_to|pos] use/[POS_VB|Stm_use|pos].

The POS tag shows this word is adjective, noun, or verb. The Stm tag is for separating the words from sentence. If the word is useful, pos is tagged in the end. In this sentence, pos = +4 because +2 for good and +2 for easy to use, neg = 0, result=(4*100)/(4+0+1)=80%. The score of the sentence is 80% after calculating the score of positive and negative words.

The negative side to this method is that it is not able to manage wide ranging opinions from users. It is necessary for the data need to do pro-process in the beginning because this allows the sentiment analysis system to make better judgments about useful opinions and if they are positive or negative.

3. CLUSTERING AND SENTIMENT ANALYSIS

3.1 Problem Report

Feature-based sentiment analysis system already upgrades word-level and sentence-level to text-level. It is acceptable to use this in the product review on Amazon because people focus on what their experience after using the products when they post product review. When we look at Twitter, people do not only talk about the experience of using product, but also many different things. The tweets from Twitter are very noisy and more spread out than the product review from Amazon. Therefore, we need to use clustering to separate all tweets into clusters to check how people think about some features of products. It can make the approach more accurate and better fit to Twitter.

3.2 Project Objective

This project objective is about receiving high accuracy sentiment analysis. First, Twitter API is processed to collect the content that includes popular electronic product name from Twitter and save to text document. In this paper, iPhone 6, Play Station 4, and Xbox One are chosen to be study cases. Second, the clustering is used to pre-process the text document and separate all tweets to some clusters. Each clusters has similar sentences or words. Third, each cluster is chosen to process in the feature-based sentiment analysis system to see the score for each cluster. Fourth, total tweets also process in the feature-based sentiment analysis system.

3.3 The Steps of the Project

Sentiment analysis has become a popular method to use for opinion mining on social networks. Generally, this method is good enough to do the job. However, the opinions on Twitter are complicated and as a result, the use of clustering is needed to organized tweets into clusters that have similarities. Twitter API, twitter4j, is used to get the tweets and save to text document [16]. K-means is chosen to do clustering to see what people’s thinking is in different features of the products. Each cluster has a high relationship and similar sentences are entered into feature-based sentiment analysis system. In addition, total tweets also process in feature-based sentiment analysis system. Before being able to run k-means on a series of text documents, the documents must be signified as equally similar directions. To accomplish this, the documents can process the TF-IDF score.

3.3.1 TF-IDF

The TF-IDF is short for term frequency-inverse document frequency. The main idea of TF-IDF is this: If a word or phrase in an article appearing in the high frequency TF, and rarely appears in other articles, you think this word or phrase has a good ability to distinguish between categories [17].

TF: the term frequency means how many times a term occurs in a document. We can calculate the term frequency for a word as the ratio of number of times the word occurs in the document to the total number of words in the document.

IDF: the inverse document frequency is a way to measure if the term is common or not for all documents. It is taken by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient [18].

The Figure 3.1 shows how to calculate TF and IDF. First, the calculation is highest when t occurs many times within a small number of documents. Second, the calculation is lower when the term occurs fewer times in a document, or occurs in multiple documents. Third, the calculation is lowest when the term occurs in almost all documents [19].

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \text{x } \text{idf}_{t}$$

Figure 3.1. The TF * IDF of Term t in Document d is Calculated

3.3.2 K-means Algorithm

K-means algorithm has some steps. First, choose k, the number of clusters to be determined. Second, choose k objects randomly as the initial cluster center. Third, assign the distance of each object to their closest cluster. We need to repeat the first and second steps couple times until no changes on cluster centers.

3.3.3 Sentiment Analysis System

Figure 3.3 demonstrates the project steps. Some clusters are gotten, and each cluster has similar sentences. Then each cluster is putted into the sentiment analysis system to find out how people think about some features of the product. In addition, total tweets also process in the sentiment analysis system. Sentiment analysis system has five steps. First, POS tagging is the method of deciding if the word is verb, adjective, or noun. Second, SentiWordNet is used for word-level opinion tagging. Third, enriching tags is for increasing or decreasing the score of the positive or negative. For example, “very good” is stronger than “good”. Fourth, sentence-level opinion mining calculates all positive and negative scores in the sentence. Fifth, document-level opinion mining is similar to sentence-level opinion mining, but at the document-level it calculates the score of all documents.

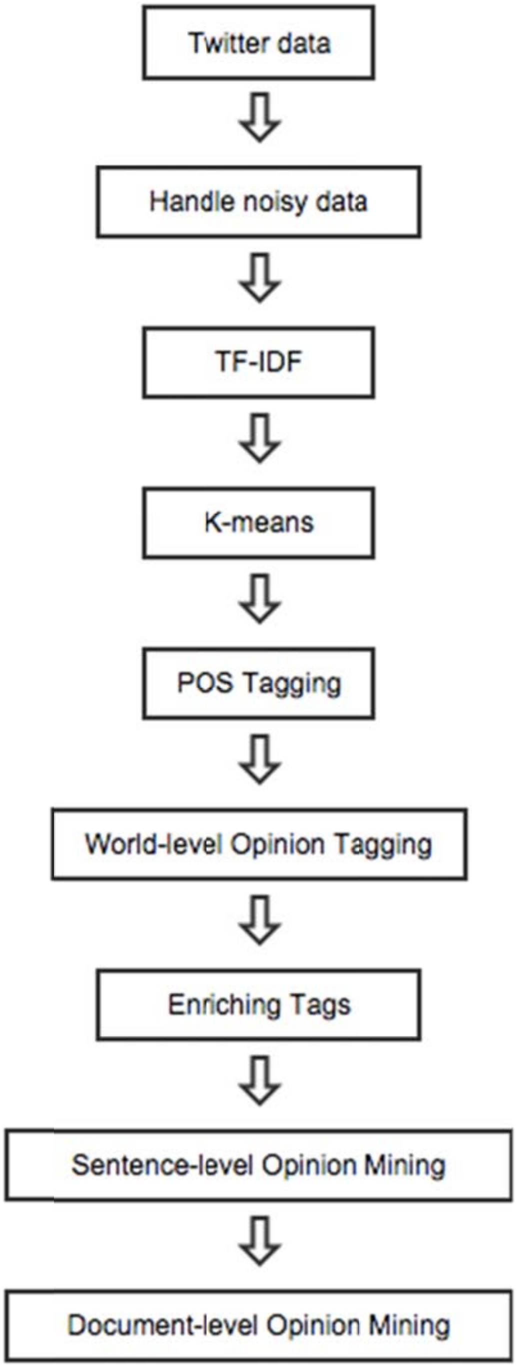

Twitter data
Handle noisy data
TF-IDF
K-means
POS Tagging
World-level Opinion Tagging
Enriching Tags
Sentence-level Opinion Mining
Document-level Opinion Mining

Figure 3.2. Project Steps

4. IMPLEMENTATION AND RESULTS

4.1 Environment

The suggested system is executed in C# and Java. For this, Java Swing and Twitter4j parser are the main programs utilized. Microsoft Visual C# and Netbeans IDE, are the programming environments used because they are more suitable for programming.

4.1.1 Microsoft Visual C#

Microsoft Visual C# is Microsoft's implementation of the C# specification, and is part of the Microsoft Visual Studio product suite [20]. C# was created by Microsoft and is a multi-paradigm programming language covering many different programming subjects, including strong typing, imperative, declarative, functional, generic, object-oriented, and component-oriented programming disciplines. [21]

4.1.2 Java Swing

Java Swing, which was released by Oracle, is a Graphical User Interface (GUI) toolkit [22]. This program lets programmers make GUI for java applications. It is stated that the parts are not heavy because of a high flexibility. Swing offers many a lot of innovative components including lists, tables, scroll panes and tabbed panels. Furthermore, there are more familiar components offered, which include labels, checkboxes and buttons. In addition, some of its components have drag and drop features to allow for further ease of use.

4.1.3 Twitter4j

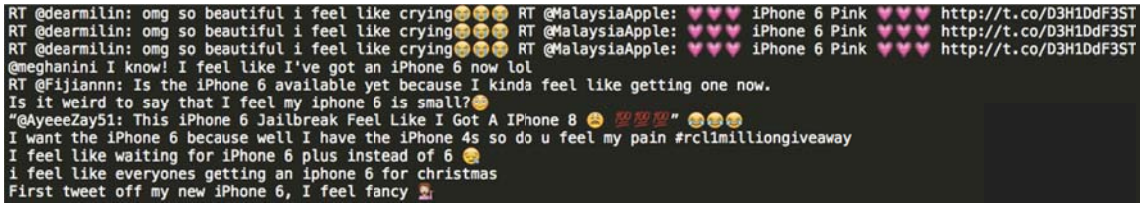
Twitter4J is an unofficial Java library for the Twitter API. With Twitter4J, you can easily integrate your Java application with the Twitter service.

4.1.4 NetBeans IDE

NetBeans is an integrated development environment (IDE) that is used mainly with Java, but it is also used with other languages, such as PHP, C/C++, and HTML5 [23]. Additionally, NetBeans is an application platform framework for not only Java desktop applications but others as well. The NetBeans IDE is written in Java and can run on Windows, OS X, Linux, Solaris and other platforms supporting a compatible JVM.

4.2 Software Modules

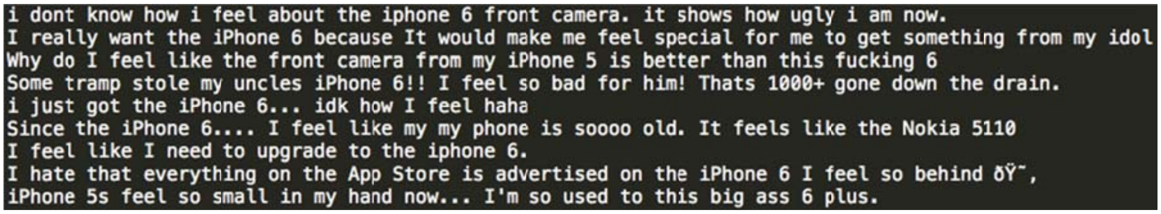
For this module, Twitter4j is used to collect the tweets. The important aspect is the text, so the user name, location and time are all ignored. Figure 4.1 shows the results of this process.



```
RT @dearmilin: omg so beautiful i feel like crying😭😭😭 RT @MalaysiaApple: 💗💗💗 iPhone 6 Pink 💗💗💗 http://t.co/D3H1DdF3ST
RT @dearmilin: omg so beautiful i feel like crying😭😭😭 RT @MalaysiaApple: 💗💗💗 iPhone 6 Pink 💗💗💗 http://t.co/D3H1DdF3ST
RT @dearmilin: omg so beautiful i feel like crying😭😭😭 RT @MalaysiaApple: 💗💗💗 iPhone 6 Pink 💗💗💗 http://t.co/D3H1DdF3ST
@meghanini I know! I feel like I've got an iPhone 6 now lol
RT @Fijiannn: Is the iPhone 6 available yet because I kinda feel like getting one now.
Is it weird to say that I feel my iphone 6 is small?😳
“@AyeeeZay51: This iPhone 6 Jailbreak Feel Like I Got A IPhone 8 😩 💯💯💯” 😂😂😂
I want the iPhone 6 because well I have the iPhone 4s so do u feel my pain #rcl1milliongiveaway
I feel like waiting for iPhone 6 plus instead of 6 😔
i feel like everyones getting an iphone 6 for christmas
First tweet off my new iPhone 6, I feel fancy 💁
```

Figure 4.1. Twitter4j Output

Unfortunately, there are a lot of noisy tweets from Twitter, so it is beneficial to use a combination of computer and human inspection to sort through the noisy tweets. The noisy tweets are checked manually to identify and eliminate outliers. The tweets include #HashName, @UserName and website link are deleted. Figure 4.2 displays the tweets after human inspection.

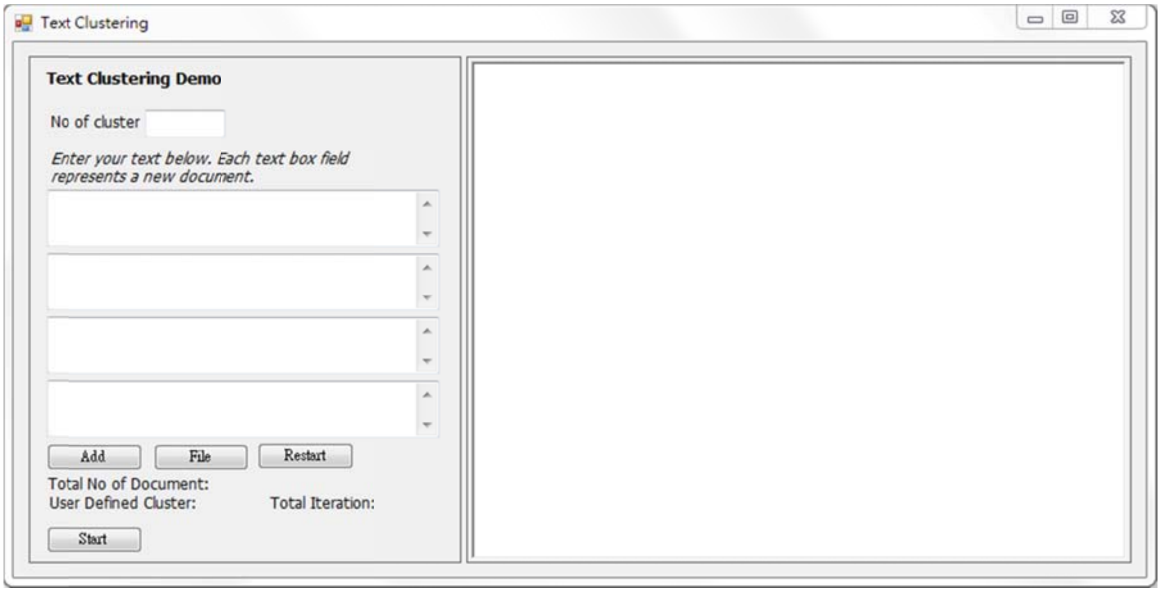


```
i dont know how i feel about the iphone 6 front camera. it shows how ugly i am now.
I really want the iPhone 6 because It would make me feel special for me to get something from my idol
Why do I feel like the front camera from my iPhone 5 is better than this fucking 6
Some tramp stole my uncles iPhone 6!! I feel so bad for him! Thats 1000+ gone down the drain.
i just got the iPhone 6... idk how I feel haha
Since the iPhone 6.... I feel like my my phone is soooo old. It feels like the Nokia 5110
I feel like I need to upgrade to the iphone 6.
I hate that everything on the App Store is advertised on the iPhone 6 I feel so behind ðŸ˜,
iPhone 5s feel so small in my hand now... I'm so used to this big ass 6 plus.
```

Figure 4.2. Tweets after Human Inspection

The interface for clustering uses C# and can be seen in Figure 4.3. At the beginning, the number of clusters must be chosen. Then the text has two ways to be entered into interface. First way is entering the text in each text box field represents a new document. The next step is to click the Add button once the text is entered. Then click the Start button after all text has been entered. If these steps are followed the then the clustering results appear on the right side.

Another way to enter tweets is from text document. Click file button to choose the text document. Then click add button to enter the data from text document. After enter the data, click start button.



Enter your text below. Each text box field represents a new document.

Figure 4.3. Clustering Interface

Figure 4.4 illustrates the User-Interface module and input handler. To complete this, first enter the text in the text space above the slider bar. The text space under the slider bar displays the sentence-level opinion mining output and the slider bar displays the entire document-level opinion mining output.

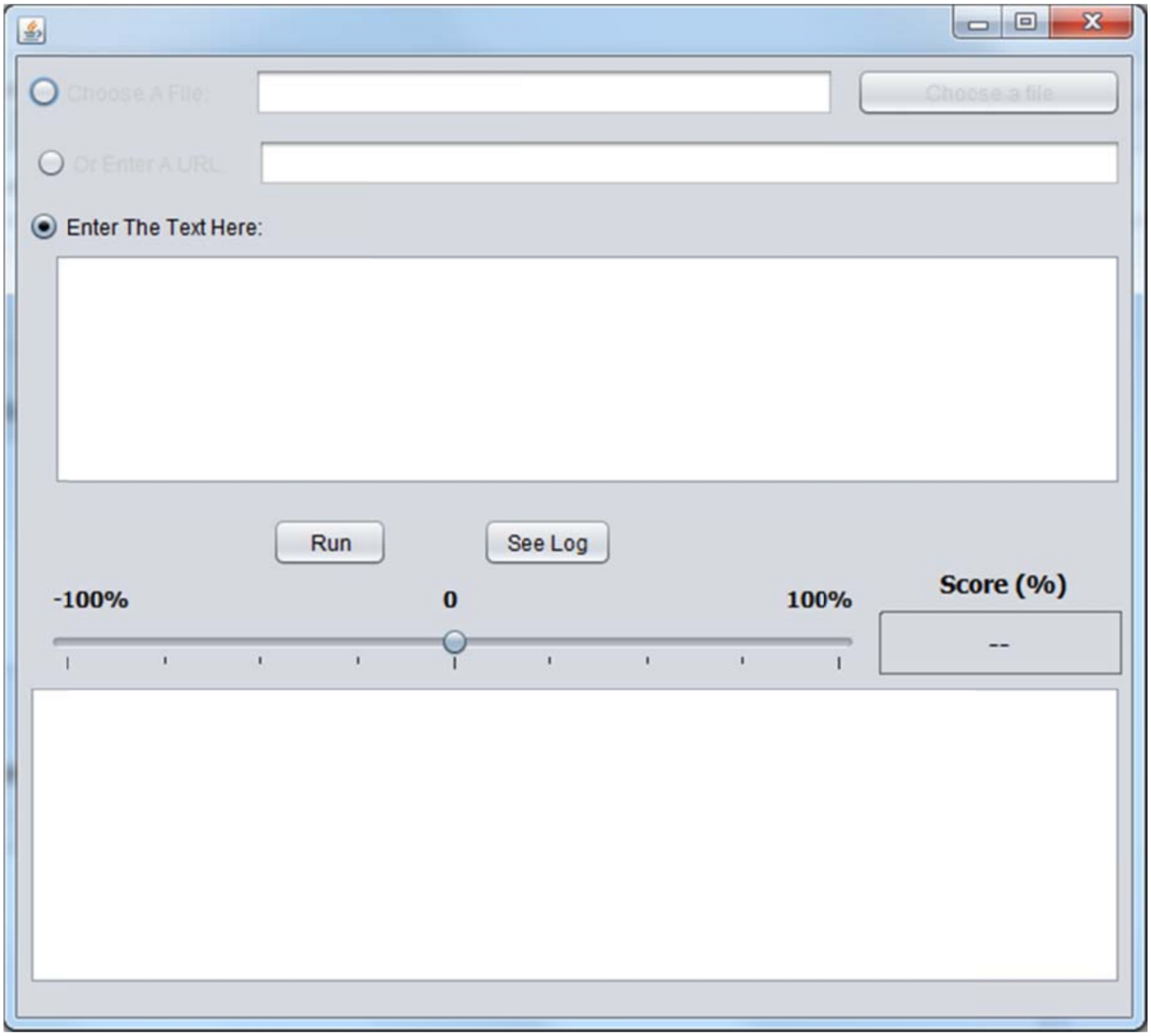
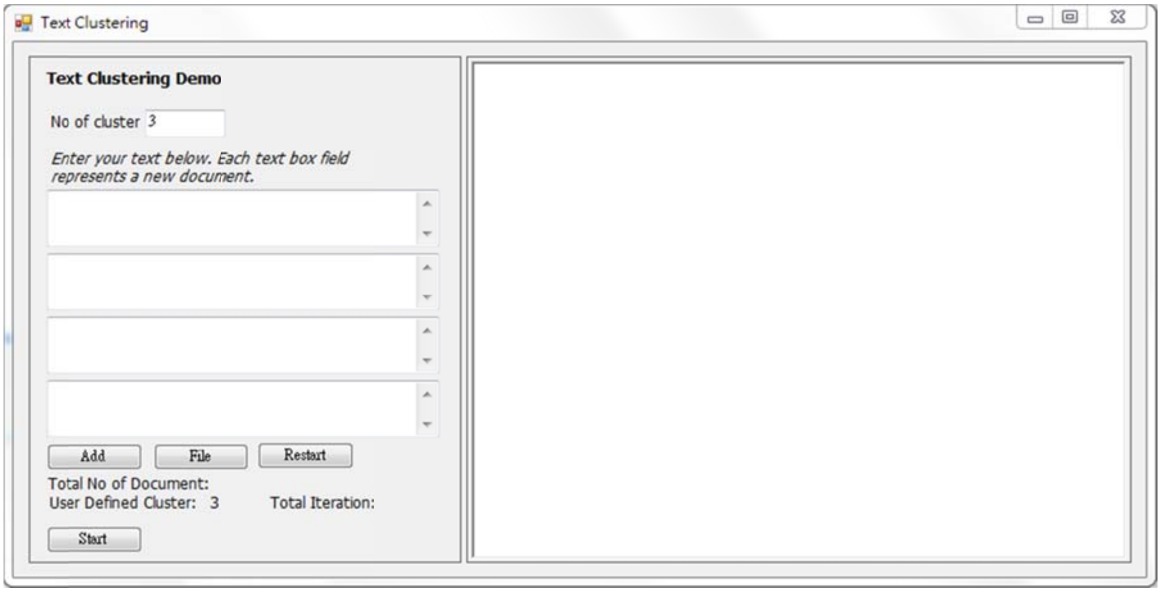


Figure 4.4. Sentiment Analysis Interface

4.3 Clustering Tweets

Figure 4.5 illustrates enter 3 to the number of cluster.



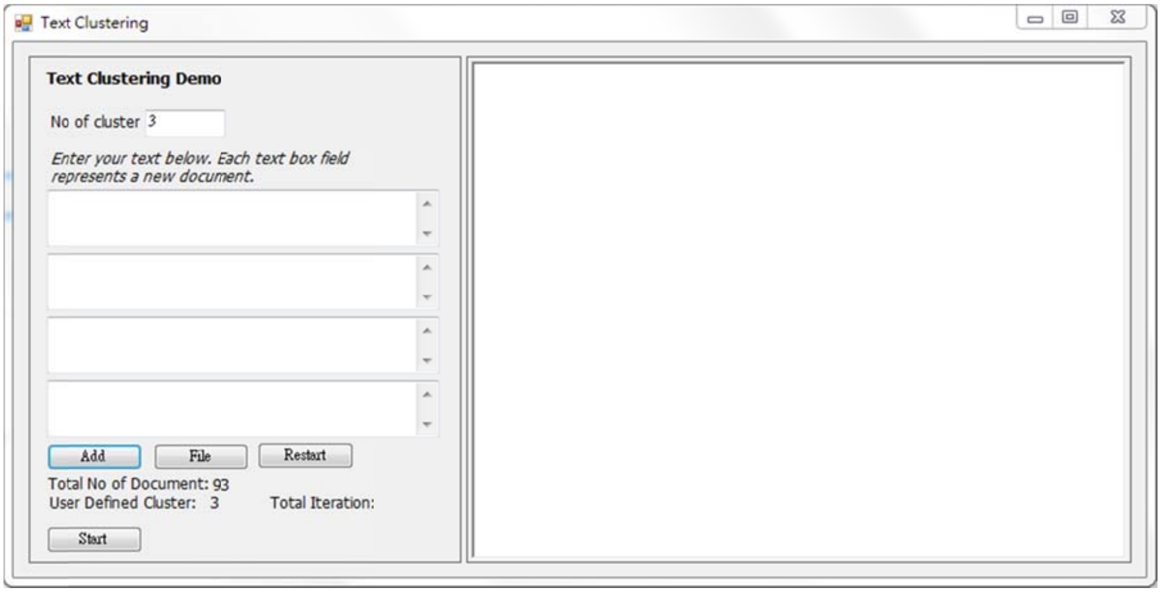
No of cluster 3

Enter your text below. Each text box field represents a new document.

Total No of Document:
User Defined Cluster: 3
Total Iteration:

Figure 4.5. Cluster Interface: Enter Cluster Number

Figure 4.6 displays click file button to choose the text document. After that, click add button to add the tweets from the text document to the clustering.



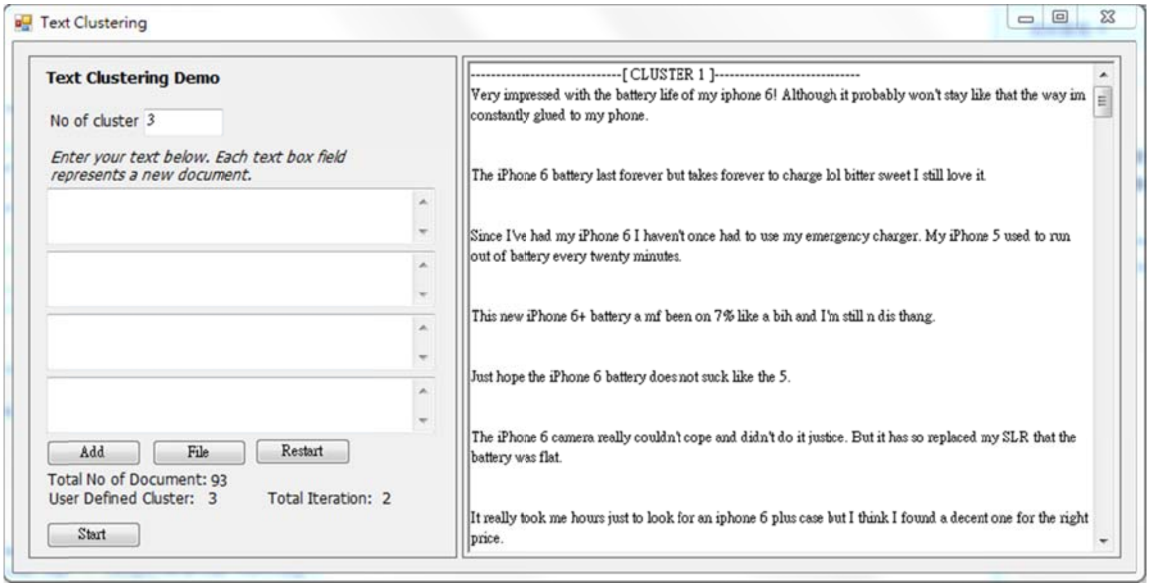
No of cluster 3

Enter your text below. Each text box field represents a new document.

Total No of Document: 93
User Defined Cluster: 3
Total Iteration:

Figure 4.6. Cluster Interface: Enter Text Document

Figure 4.7 displays the cluster 1 once all tweets are entered and the clustering is completed.


Text Clustering
Text Clustering Demo
Enter your text below. Each text box field represents a new document.
Add
File
Restart
Start

No of cluster 3

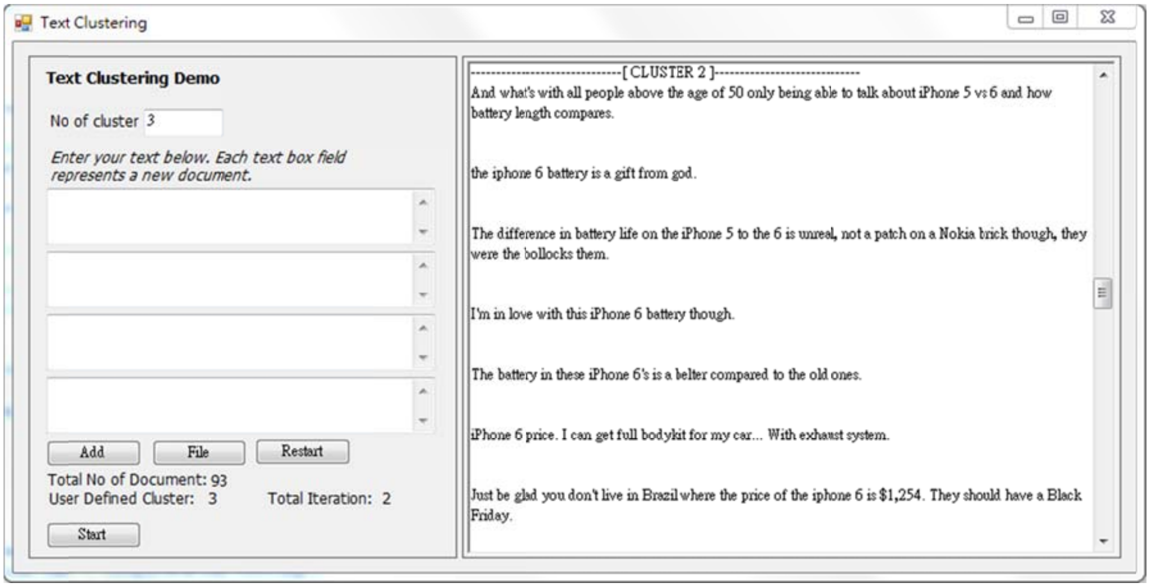
Total No of Document: 93

User Defined Cluster: 3 Total Iteration: 2

```
------------------------------[ CLUSTER 1 ]----------------------------
Very impressed with the battery life of my iphone 6! Although it probably won't stay like that the way im
constantly glued to my phone.


The iPhone 6 battery last forever but takes forever to charge lol bitter sweet I still love it


Since I've had my iPhone 6 I haven't once had to use my emergency charger. My iPhone 5 used to run
out of battery every twenty minutes.


This new iPhone 6+ battery a mf been on 7% like a bih and I'm still n dis thang.


Just hope the iPhone 6 battery does not suck like the 5.


The iPhone 6 camera really couldn't cope and didn't do it justice. But it has so replaced my SLR that the
battery was flat.


It really took me hours just to look for an iphone 6 plus case but I think I found a decent one for the right
price.
```

Figure 4.7. Cluster 1

Figure 4.8 shows the cluster 2 once all data is entered and the clustering is completed.


Text Clustering
Text Clustering Demo
Enter your text below. Each text box field represents a new document.
Add
File
Restart
Start

No of cluster 3

Total No of Document: 93

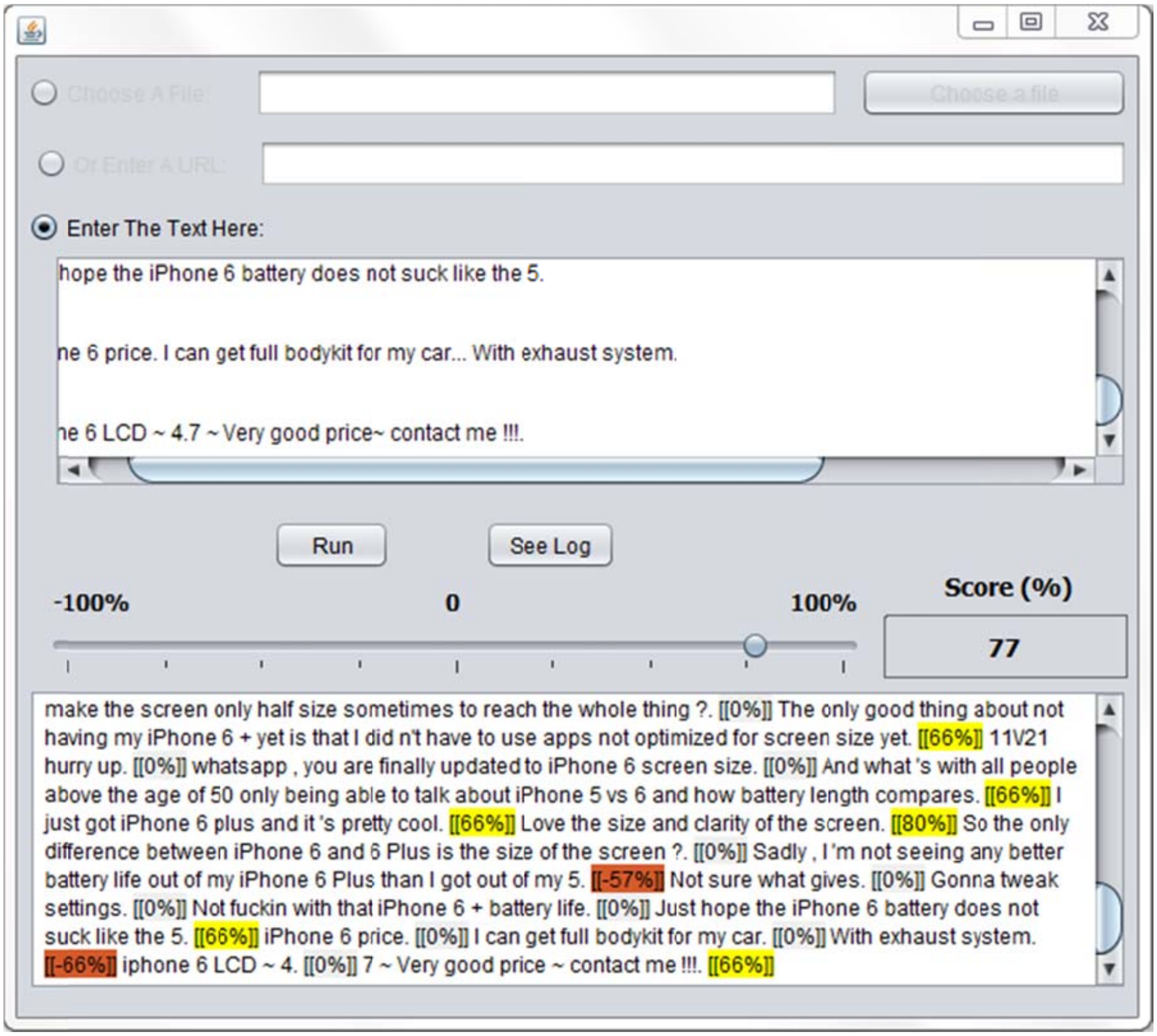
User Defined Cluster: 3 Total Iteration: 2

```
------------------------------[ CLUSTER 2 ]----------------------------
And what's with all people above the age of 50 only being able to talk about iPhone 5 vs 6 and how
battery length compares.


the iphone 6 battery is a gift from god.


The difference in battery life on the iPhone 5 to the 6 is unreal, not a patch on a Nokia brick though, they
were the bollocks them.


I'm in love with this iPhone 6 battery though.


The battery in these iPhone 6's is a belter compared to the old ones.


iPhone 6 price. I can get full bodykit for my car... With exhaust system.


Just be glad you don't live in Brazil where the price of the iphone 6 is $1,254. They should have a Black
Friday.
```

Figure 4.8. Cluster 2

4.4 Sentiment Analysis

Figure 4.9 shows how the cluster 1 is selected and how that tweets are inputted into the sentiment analysis to receive a score. The range of score is from 100% to -100%. 100% means the most positive opinion. -100% means the most negative opinion. The score of each sentence shows in the end of the sentence. After that, the system adds all scores together and outputs the final score.



○ Choose A File:

○ Or Enter A URL:

● Enter The Text Here:

```
hope the iPhone 6 battery does not suck like the 5.

ne 6 price. I can get full bodykit for my car... With exhaust system.

ne 6 LCD ~ 4.7 ~ Very good price~ contact me !!!.
```

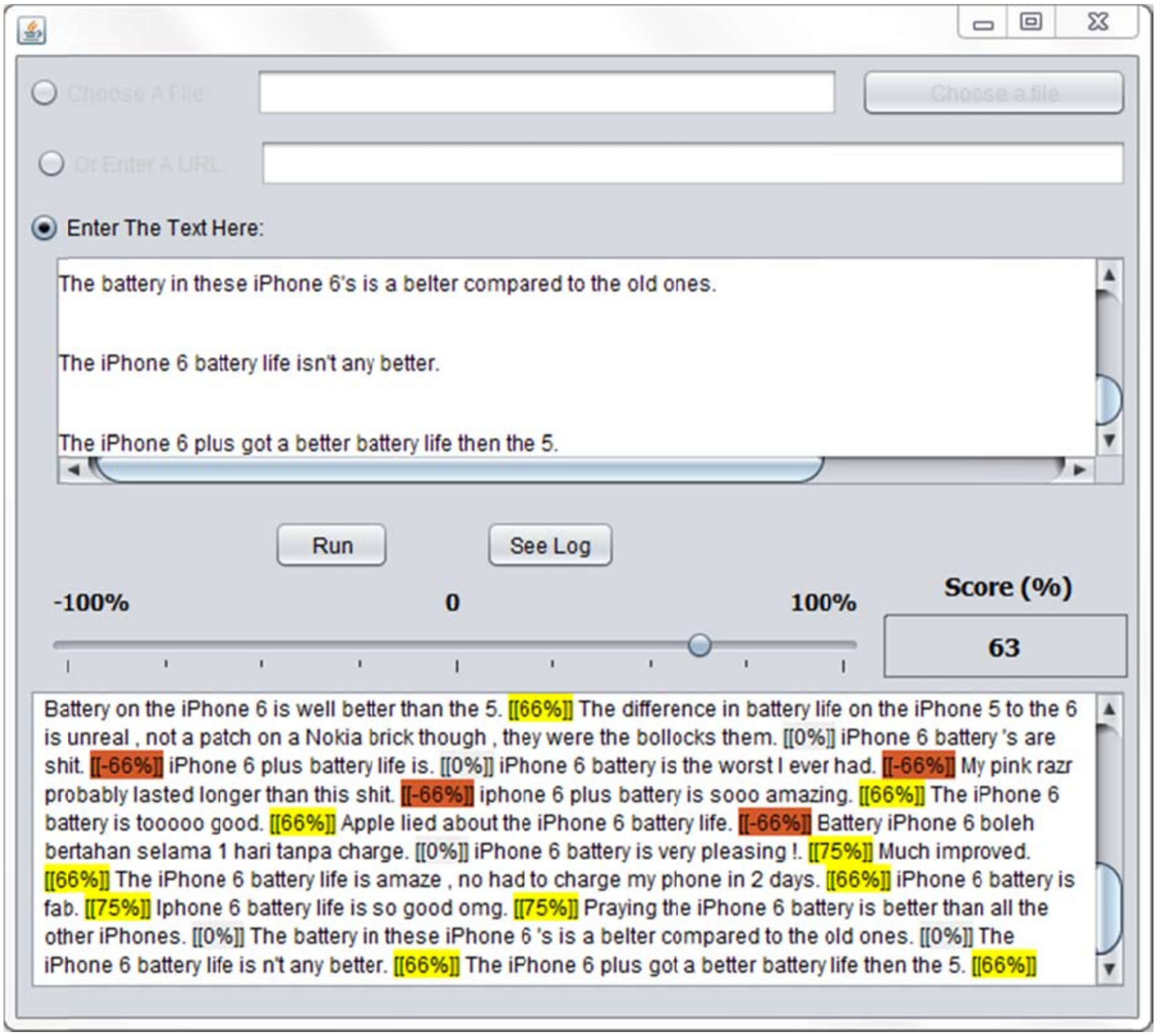
Score (%)

77

```
make the screen only half size sometimes to reach the whole thing ?. [[0%]] The only good thing about not
having my iPhone 6 + yet is that I did n't have to use apps not optimized for screen size yet. [[66%]] 11V21
hurry up. [[0%]] whatsapp , you are finally updated to iPhone 6 screen size. [[0%]] And what 's with all people
above the age of 50 only being able to talk about iPhone 5 vs 6 and how battery length compares. [[66%]] I
just got iPhone 6 plus and it 's pretty cool. [[66%]] Love the size and clarity of the screen. [[80%]] So the only
difference between iPhone 6 and 6 Plus is the size of the screen ?. [[0%]] Sadly , I 'm not seeing any better
battery life out of my iPhone 6 Plus than I got out of my 5. [[-57%]] Not sure what gives. [[0%]] Gonna tweak
settings. [[0%]] Not fuckin with that iPhone 6 + battery life. [[0%]] Just hope the iPhone 6 battery does not
suck like the 5. [[66%]] iPhone 6 price. [[0%]] I can get full bodykit for my car. [[0%]] With exhaust system.
[[-66%]] iphone 6 LCD ~ 4. [[0%]] 7 ~ Very good price ~ contact me !!!. [[66%]]
```

Figure 4.9. Sentiment Analysis: Score of the Cluster 1

Figure 4.10 illustrates how the cluster 2 is selected and how that tweets are inputted into the sentiment analysis to receive a score



● Enter The Text Here:

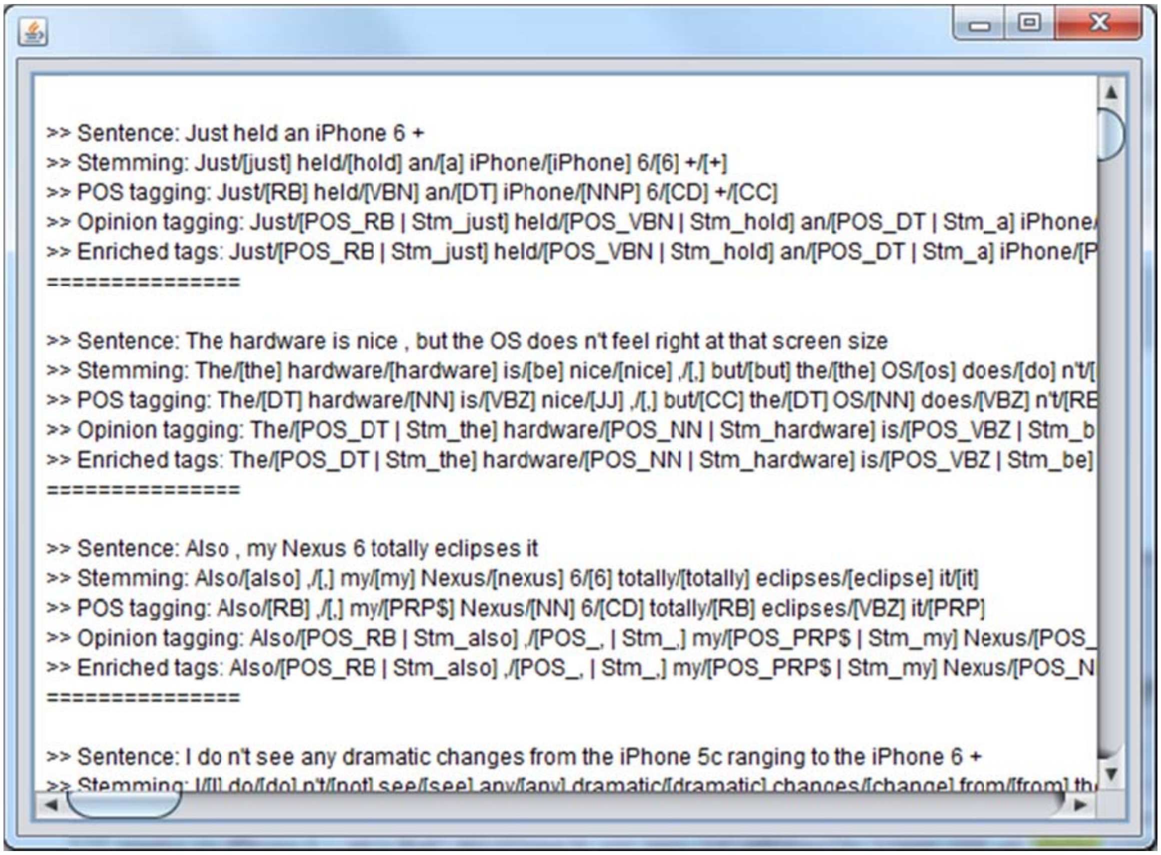
```
The battery in these iPhone 6's is a belter compared to the old ones.

The iPhone 6 battery life isn't any better.

The iPhone 6 plus got a better battery life then the 5.
```

Score (%): 63

Battery on the iPhone 6 is well better than the 5. [[66%]] The difference in battery life on the iPhone 5 to the 6 is unreal , not a patch on a Nokia brick though , they were the bollocks them. [[0%]] iPhone 6 battery 's are shit. [[-66%]] iPhone 6 plus battery life is. [[0%]] iPhone 6 battery is the worst I ever had. [[-66%]] My pink razr probably lasted longer than this shit. [[-66%]] iphone 6 plus battery is sooo amazing. [[66%]] The iPhone 6 battery is tooooo good. [[66%]] Apple lied about the iPhone 6 battery life. [[-66%]] Battery iPhone 6 boleh bertahan selama 1 hari tanpa charge. [[0%]] iPhone 6 battery is very pleasing !. [[75%]] Much improved. [[66%]] The iPhone 6 battery life is amaze , no had to charge my phone in 2 days. [[66%]] iPhone 6 battery is fab. [[75%]] Iphone 6 battery life is so good omg. [[75%]] Praying the iPhone 6 battery is better than all the other iPhones. [[0%]] The battery in these iPhone 6 's is a belter compared to the old ones. [[0%]] The iPhone 6 battery life is n't any better. [[66%]] The iPhone 6 plus got a better battery life then the 5. [[66%]]

Figure 4.10. Sentiment Analysis: Score of the Cluster 2

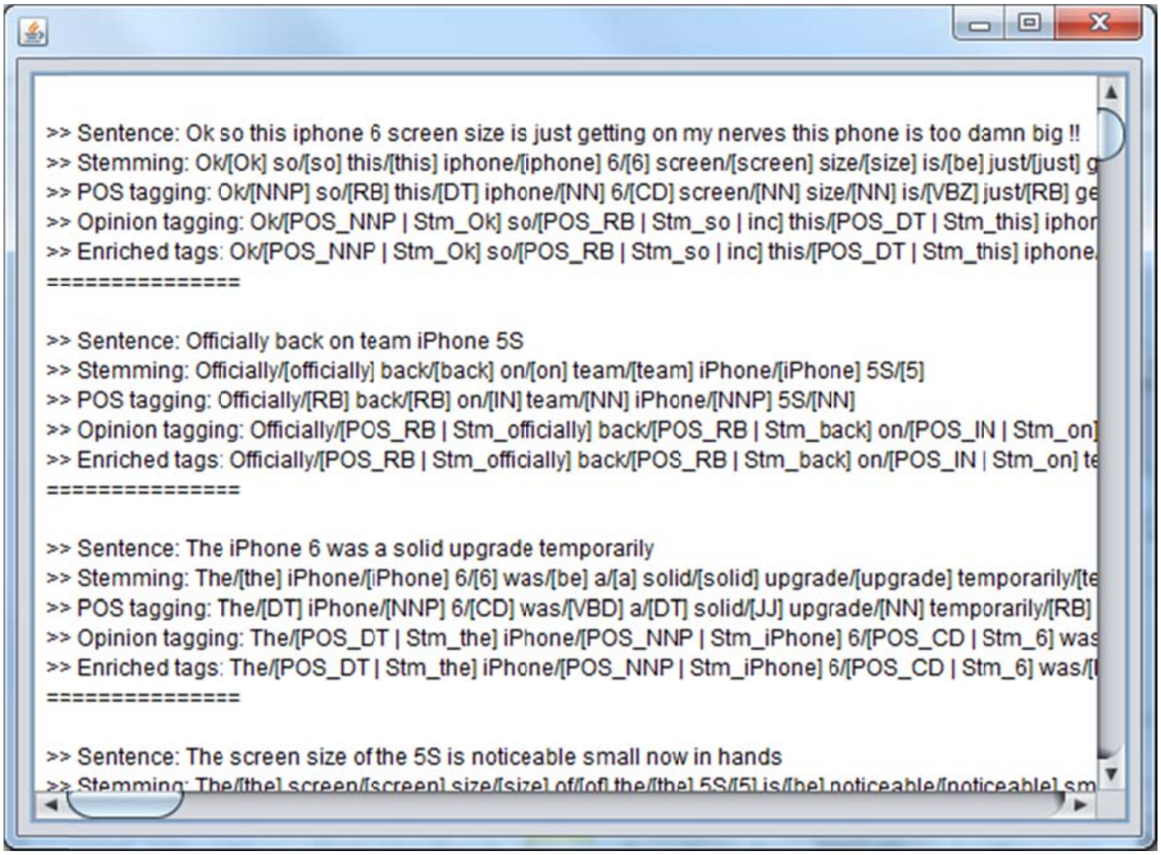
Figure 4.11 illustrates stemming, POS tagging, word-level opinion tagging and enriching tags. For example, the POS tagging of the sentence, “Just held an iPhone6 +”, is “Just/[RB] held/[VBN] an/[DT] iPhone/[NNP] 6/[CD] +/[CC]”.



```
>> Sentence: Just held an iPhone 6 +
>> Stemming: Just/[just] held/[hold] an/[a] iPhone/[iPhone] 6/[6] +/[+]
>> POS tagging: Just/[RB] held/[VBN] an/[DT] iPhone/[NNP] 6/[CD] +/[CC]
>> Opinion tagging: Just/[POS_RB | Stm_just] held/[POS_VBN | Stm_hold] an/[POS_DT | Stm_a] iPhone/
>> Enriched tags: Just/[POS_RB | Stm_just] held/[POS_VBN | Stm_hold] an/[POS_DT | Stm_a] iPhone/[P
================

>> Sentence: The hardware is nice , but the OS does n't feel right at that screen size
>> Stemming: The/[the] hardware/[hardware] is/[be] nice/[nice] ,/[,] but/[but] the/[the] OS/[os] does/[do] n't/[
>> POS tagging: The/[DT] hardware/[NN] is/[VBZ] nice/[JJ] ,/[,] but/[CC] the/[DT] OS/[NN] does/[VBZ] n't/[RB
>> Opinion tagging: The/[POS_DT | Stm_the] hardware/[POS_NN | Stm_hardware] is/[POS_VBZ | Stm_b
>> Enriched tags: The/[POS_DT | Stm_the] hardware/[POS_NN | Stm_hardware] is/[POS_VBZ | Stm_be]
================

>> Sentence: Also , my Nexus 6 totally eclipses it
>> Stemming: Also/[also] ,/[,] my/[my] Nexus/[nexus] 6/[6] totally/[totally] eclipses/[eclipse] it/[it]
>> POS tagging: Also/[RB] ,/[,] my/[PRP$] Nexus/[NN] 6/[CD] totally/[RB] eclipses/[VBZ] it/[PRP]
>> Opinion tagging: Also/[POS_RB | Stm_also] ,/[POS_, | Stm_,] my/[POS_PRP$ | Stm_my] Nexus/[POS_
>> Enriched tags: Also/[POS_RB | Stm_also] ,/[POS_, | Stm_,] my/[POS_PRP$ | Stm_my] Nexus/[POS_N
================

>> Sentence: I do n't see any dramatic changes from the iPhone 5c ranging to the iPhone 6 +
>> Stemming: I/[I] do/[do] n't/[not] see/[see] any/[any] dramatic/[dramatic] changes/[change] from/[from] th
```

Figure 4.11. Sentiment Analysis: Tagging of the Cluster 1

Figure 4.12 shows stemming, POS tagging, word-level opinion tagging and enriching tags.



```
>> Sentence: Ok so this iphone 6 screen size is just getting on my nerves this phone is too damn big !!
>> Stemming: Ok/[Ok] so/[so] this/[this] iphone/[iphone] 6/[6] screen/[screen] size/[size] is/[be] just/[just] g
>> POS tagging: Ok/[NNP] so/[RB] this/[DT] iphone/[NN] 6/[CD] screen/[NN] size/[NN] is/[VBZ] just/[RB] ge
>> Opinion tagging: Ok/[POS_NNP | Stm_Ok] so/[POS_RB | Stm_so | inc] this/[POS_DT | Stm_this] iphor
>> Enriched tags: Ok/[POS_NNP | Stm_Ok] so/[POS_RB | Stm_so | inc] this/[POS_DT | Stm_this] iphone/
================

>> Sentence: Officially back on team iPhone 5S
>> Stemming: Officially/[officially] back/[back] on/[on] team/[team] iPhone/[iPhone] 5S/[5]
>> POS tagging: Officially/[RB] back/[RB] on/[IN] team/[NN] iPhone/[NNP] 5S/[NN]
>> Opinion tagging: Officially/[POS_RB | Stm_officially] back/[POS_RB | Stm_back] on/[POS_IN | Stm_on]
>> Enriched tags: Officially/[POS_RB | Stm_officially] back/[POS_RB | Stm_back] on/[POS_IN | Stm_on] te
================

>> Sentence: The iPhone 6 was a solid upgrade temporarily
>> Stemming: The/[the] iPhone/[iPhone] 6/[6] was/[be] a/[a] solid/[solid] upgrade/[upgrade] temporarily/[te
>> POS tagging: The/[DT] iPhone/[NNP] 6/[CD] was/[VBD] a/[DT] solid/[JJ] upgrade/[NN] temporarily/[RB]
>> Opinion tagging: The/[POS_DT | Stm_the] iPhone/[POS_NNP | Stm_iPhone] 6/[POS_CD | Stm_6] was
>> Enriched tags: The/[POS_DT | Stm_the] iPhone/[POS_NNP | Stm_iPhone] 6/[POS_CD | Stm_6] was/[
================

>> Sentence: The screen size of the 5S is noticeable small now in hands
>> Stemming: The/[the] screen/[screen] size/[size] of/[of] the/[the] 5S/[5] is/[be] noticeable/[noticeable] sm
```

Figure 4.12. Sentiment Analysis: Tagging of the Cluster 2

5. TESTING AND EVALUATION

iPhone 6, Play Station 4, and Xbox One were chosen as keywords to search on Twitter. Tweets with these keywords were collected and saved to the text document. Once the tweets are collected, the clustering is done followed by processing the sentiment analysis system. This is because the tweets relative to different features of products. At the time of clustering, the k-means algorithm is used to deal with the tweets, and k is set to 3.

5.1 iPhone 6

In the iPhone 6, after human inspection, the data set has a total of 88 tweets. Once the clustering is processed, 3 clusters are taken. Cluster 1 has 31 tweets, cluster 2 has 37 tweets, and cluster 3 has 20 tweets. The clusters are added into the sentiment analysis system in order to compute the score. Table 5.1 shows the result of this computation.

Table 5.1. iPhone 6 Clusters and Score

Cluster	Tweets	Score (%)	Feature
1	31	77	screen
2	37	63	battery
3	20	71	price
Total	88	71	

Cluster 1 contains 80.6% tweets relative to screen size (25 out of 31 tweets). Cluster 2 has 86.5% tweets relative to battery life (32 out of 37 tweets). Cluster 3 includes 85% tweets that mentioned price (17 out of 20 tweets). People are more satisfied with the iPhone 6 screen size compared with the battery life by looking at the scores. The score of the iPhone 6 screen size is 77%, and the score of the battery life is only 63%.

A few people are asked to manually judge if this content is positive or negative. After that, classifier evaluation metrics and confusion matrix are used to check the score from this project and the judgment from the people who review the content [24].

Table 5.2 shows the evaluation report of iPhone 6. True positives (TP) means human’s check and system output are both positive. True negative (FP) means human’s check and system output are both negative. TP and FP mean the system output has correct determine. False negative (FN) means human’s check is positive, but system output is negative. False positive (FP) means human’s check is negative, but system output is positive. FN and FP means the system output has wrong determine. ~FN and ~FP means the tweets are not about positive and negative.

Table 5.2. Evaluation Report of iPhone 6

Manual(human)/System Output	Positive (Score > 0%)	Neutral (Score = 0%)	Negative (Score < 0%)
Positive	42 (TP)	15 (~FN)	2 (FN)
Negative	3 (FP)	14 (~FP)	12 (TN)

Accuracy of this system developed means percentage of test set tuples that are correctly classified. It is calculated by using the following formula.

Opinion Extraction Accuracy = (TP+TN)/(TP+TN+FP+FN)

= (42 + 12) / (42 + 12 + 3 + 2)

= 91.5 %

Precision means what % of tuples that the classifier labeled as positive is actually positive. It is calculated by using the following formulas.

Precision = TP/(TP+FP)

= 42 / (42 + 3)

= 93.3 %

Recall means what % of positive tuples did the classifier labeled as positive. It is calculated by using the following formulas.

Recall = TP/(TP+FN)

= 42 / (42 + 2)

= 95.5 %

5.2 Play Station 4

In Play Station 4 (PS4), data set has total of 92 tweets after human inspection. After processing clustering, 3 clusters are retrieved. Cluster 1 has 34 tweets, cluster 2 has 21 tweets, and cluster 3 has 37 tweets. Each cluster is entered into the sentiment analysis system to calculate the score. Table 5.3 shows the result.

Table 5.3. PS4 Clusters and Score

Cluster	Tweets	Score (%)	Feature
1	34	51	controller
2	21	67	game
3	37	72	price
Total	92	64	

Cluster 1 contains 82.4% tweets relative to PS4 controller (28 out of 34 tweets). Cluster 2 has 81% tweets are about PS4 game (17 out of 21 tweets). Cluster 3 includes

78.4% tweets mentioned price (29 out of 37 tweets). People are not satisfied with the PS4 controller compared with the price based on the scores. The score of the PS4 controller is just 51%, whereas the score of the price is 72%.

Table 5.4 shows the evaluation report of PS4.

Table 5.4. Evaluation Report of PS4

Manual(human)/System Output	Positive (Score > 0%)	Neutral (Score = 0%)	Negative (Score < 0%)
Positive	30 (TP)	22 (~FN)	3 (FN)
Negative	9 (FP)	11 (~FP)	17 (TN)

Opinion Extraction Accuracy = (30 + 17) / (30 + 17 + 9 + 3)

= 79.7 %

Precision = 30 / (30 + 9)

= 76.9 %

Recall = 30 / (30 + 3)

= 90.9 %

5.3 Xbox One

For Xbox One, data set has total of 109 tweets after human inspection. After processing clustering, 3 clusters are retrieved. Cluster 1 has 38 tweets, cluster 2 has 23 tweets, and cluster 3 has 48 tweets. Each cluster is entered into the sentiment analysis system to calculate the score. Table 5.5 shows the result.

Table 5.5. Xbox One Clusters and Score

Cluster	Tweets	Score(%)	Feature
1	38	60	game
2	23	-59	price
3	48	55	controller
Total	109	53	

Cluster 1 contains 86.8% tweets relative to Xbox One game (33 out of 38 tweets). Cluster 2 has 78.3% tweets are about price (18 out of 23 tweets). Cluster 3 includes 79.2% tweets mentioned Xbox One controller (38 out of 48 tweets). People are not satisfied with the price of the Xbox and think it is too expensive. The score of the price is negative (-59%).

Table 5.6 shows the evaluation report of Xbox One.

Table 5.6. Evaluation Report of Xbox One

Manual(human)/System Output	Positive (Score > 0%)	Neutral (Score = 0%)	Negative (Score < 0%)
Positive	27 (TP)	22 (~FN)	10 (FN)
Negative	4 (FP)	17 (~FP)	29 (TN)

Opinion Extraction Accuracy = (27 + 29) / (27 + 29 + 4 + 10)

= 80 %

Precision = 27 / (27 + 4)

= 87.1 %

Recall = 27 / (27 + 10)

= 73 %

Table 5.7. Compare PS4 and Xbox One

	PS4 score(%)	Xbox one score(%)
game	67	60
price	72	-59
controller	51	55
total	64	53

Table 5.7 shows a comparison of the PS4 and Xbox One. In the game, people are more satisfied with the PS4 game than the Xbox One game. In the price, most people think the price of the PS4 is fine (72%), but they think the price of the Xbox One is too expensive (-59%). In the controller, people like the Xbox One controller a little more.

Actually, the PS4 has better sales than the Xbox One in USA. Figure 5.1 shows the cumulative U.S. sales since the release of Sony's PS4 and Microsoft's Xbox One.

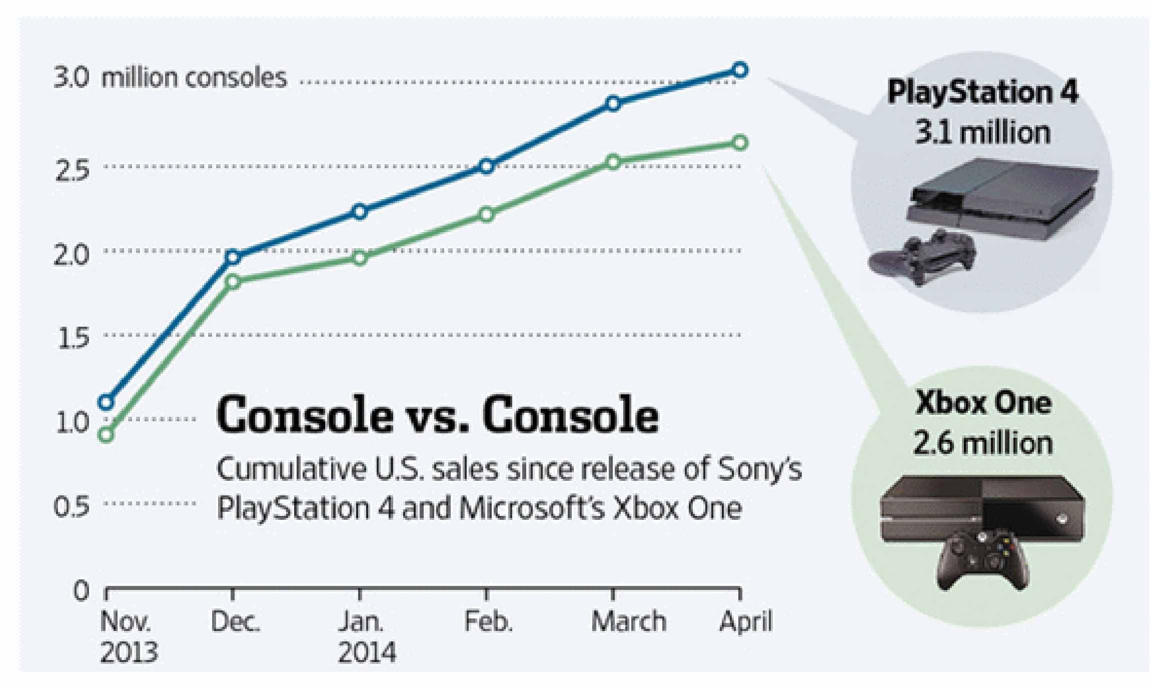
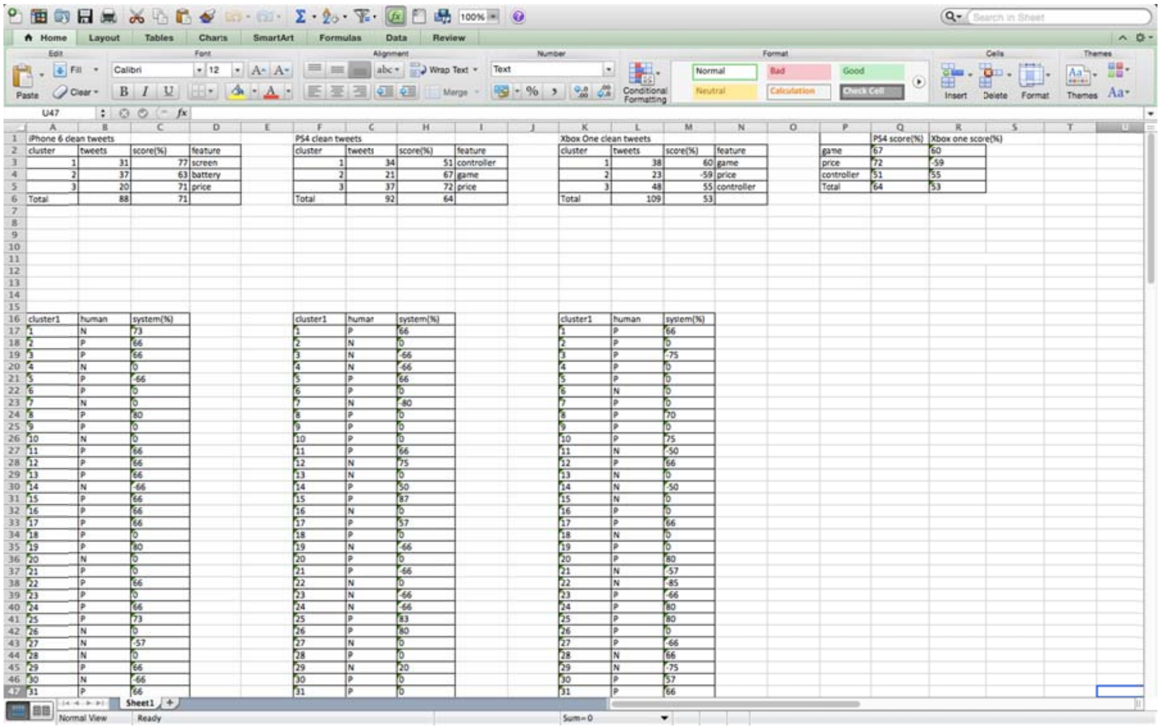


Figure 5.1. U.S. Sales of PS4 and Xbox One [25]

Figure 5.2 shows the system output for all data.



iPhone 6 clean tweets

cluster	tweets	score(%)	feature
1	31	77	screen
2	37	63	battery
3	20	71	price
Total	88	71	

PS4 clean tweets

cluster	tweets	score(%)	feature
1	34	51	controller
2	21	67	game
3	37	72	price
Total	92	64	

Xbox One clean tweets

cluster	tweets	score(%)	feature
1	38	60	game
2	23	-59	price
3	48	55	controller
Total	109	53	

	PS4 score(%)	Xbox one score(%)
game	67	60
price	72	-59
controller	51	55
Total	64	53

cluster1	human	system(%)
1	N	73
2	P	66
3	P	66
4	N	0
5	P	-66
6	P	0
7	N	0
8	P	80
9	P	0
10	N	0
11	P	66
12	P	66
13	P	66
14	N	-66
15	P	66
16	P	66
17	P	66
18	P	0
19	P	80
20	N	0
21	P	0
22	P	66
23	P	0
24	P	66
25	P	73
26	N	0
27	N	-57
28	N	0
29	P	66
30	N	-66
31	P	66

cluster1	human	system(%)
1	P	66
2	N	0
3	N	-66
4	N	-66
5	P	66
6	P	0
7	N	-80
8	P	0
9	P	0
10	P	0
11	P	66
12	N	75
13	N	0
14	P	50
15	P	87
16	N	0
17	P	57
18	P	0
19	N	-66
20	P	0
21	P	-66
22	N	0
23	N	-66
24	N	-66
25	P	83
26	P	80
27	N	0
28	P	0
29	N	20
30	P	0
31	P	0

cluster1	human	system(%)
1	P	66
2	P	0
3	P	-75
4	P	0
5	P	0
6	N	0
7	P	0
8	P	70
9	P	0
10	P	75
11	N	-50
12	P	66
13	N	0
14	N	-50
15	N	0
16	P	0
17	P	66
18	N	0
19	P	0
20	P	80
21	N	-57
22	N	-85
23	P	-66
24	P	80
25	P	80
26	P	0
27	P	-66
28	N	66
29	N	-75
30	P	57
31	P	66

Figure 5.2. System Output for All Data

6. CONCLUSION AND FUTURE WORK

This project can find how people think about specific popular electronic products. This project changes people's words to numbers and then these numbers can be analyzed to understand the different people's thinking. The problem is making sure that the change is correct. Therefore, I process the clustering and feature-based sentiment analysis system to help with the accuracy of the change.

The clustering and feature-based sentiment analysis system processes the text document from Twitter. Because the opinions on Twitter are too complex and dispersed, clustering needs to be used to separate data into clusters. In this paper, Twitter API, twitter4j, is used to get the data and save to text document. Then k-means algorithm is used to do clustering. After that, feature-based sentiment analysis system is used to process the data. The sentiment analysis system is done in seven main steps: stemming, POS tagging, word-level opinion tagging, enriching tags, sentence-level opinion mining, document-level opinion mining, and time-level opinion mining. the Stanford tool is used to process the stemming and POS tagging. Then SentiWordNet is used to handle the enriching tags and word-level tags.

Apart from the work done towards this system, future work mainly comprises of the following objectives.

- To handle the noisy data without human inspection.
- To improve the speed with a large number of sentences and handle huge data.
- To run this project on Cloud computing with Hadoop and Mahout.
- Run sentiment analysis in Chinese on Weibo.

BIBLIOGRAPHY AND REFERENCES

[1] Liu, B. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

[2] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.\

[3] Weibo. http://en.wikipedia.org/wiki/Sina_Weibo

[4] Xue, B., Fu, C., & Shaobin, Z. (2014, June). A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec. In *Big Data (BigData Congress), 2014 IEEE International Congress on* (pp. 358-363). IEEE.

[5] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc..

[6] Text Documents Clustering using K-Means Algorithm. http://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm

[7] Tushar Khot,Clustering Twitter Feeds using Word Co-occurrence CS769 Project Report. http://pages.cs.wisc.edu/~tushar/projects/cs769.pdf

[8] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Applied statistics, 100-108.

[9] Han, J., & Kamber, M. (2006). Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann.

[10] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

[11] Bora, N. N. (2011). Feature Based Sentiment Analysis on Twitter (Doctoral dissertation, Indian Institute of Technology Guwahati).

[12] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze (2008). “Introduction to Information Retrieval,” Cambridge University Press. http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[13] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani (2010). “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.”

[14] Srividya Venumbaka (Spring 2013). “An Enhanced Feature-Based Sentiment Analysis System.” Graduate Project Report. Texas A&M University Corpus Christi.

[15] The Stanford Natural Language Processing Group. (n.d.) “Stanford log-linear Part-of-Speech Tagger.” http://nlp.stanford.edu/software/tagger.shtml

[16] Twitter4J. (2013). http://twitter4j.org/en/index.html

[17] Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press.

[18] TF-IDF means. http://www.tfidf.com/

[19] The Stanford Natural Language Processing Group. TD-IDF weighting. http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

[20] Microsoft Visual C#. http://en.wikipedia.org/wiki/Microsoft_Visual_C_Sharp

[21] C#. http://en.wikipedia.org/wiki/C_Sharp_(programming_language)

[22] Java Swing. http://en.wikibooks.org/wiki/Java_Swings

[23] NetBeans IDE. http://en.wikipedia.org/wiki/NetBeans

[24] Kohavi and Provost. (1998). ConfusionMatrix.

http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

[25] Wall Street Journal. http://iknow.stpi.narl.org.tw/post/Read.aspx?PostID=9775