A Study on Customers' Sentiment Analysis Based on Big Data Using Twitter Data

Xiaorui Shao, Chang-So Kim, and Kwak Dong Ryul

***Abstract*—This paper focuses on mining the value of customers emotional behaviors using Twitter data. Using Apache Flume to collect tweets data from Twitter. 192,390 tweets were collected. Then the natural language processing (NLP) technology has been used to divide and filter tweets for customers' emotional behaviors analysis. We picked five main hot topics among these tweets. Choose one of the hot topics HUAWEI honors 9 for sentiment analysis (SA). Compared with Native Bayes, Maximum Entropy Classifier. Decision Tree Classifier is the most effective classification method for our data sets. According to our experiment, the result shows that 45% of customers are satisfied with HUAWEI honors 9, but there is still having 36% of customers unsatisfied with it. Specifically, in the field of battery, game and stand-by power consumption, it needs a great of an improvement.**

***Index Terms*—Customers' emotional behavior, twitter, big data, sentiment analysis.**

I. Introduction

Twitter is one of the most fashionable social platforms where registered users can update their messages and follow others' statuses expediently. The user can use less than 140 characters (named tweets) to express their behaviors, state of mind, comments on certain topics, etc. Tweets data sentiment can feedback much more important information we have never thought of (e.g., the trend of the stock market [1], [2]. Thus, tweets can be used to explain, detect, or predict various phenomena [3]. However, most of the research only focused on analysis, no more about connecting it with big data. In this paper, using flume to collect tweets data. Collected tweets are stored in HDFS. Using python imports tweets to analyze. Through analyzing massive tweets, we get more detailed information.

II. Work Flow

The whole workflow is divided into four parts: Data collection, Preprocessing, sentiment analysis, and conclusions. The workflow as shown in Fig. 1.

Data collection: Using Apache flume which is one of the components in Hadoop to collect tweets about HUAWEI Smartphones by keywords.

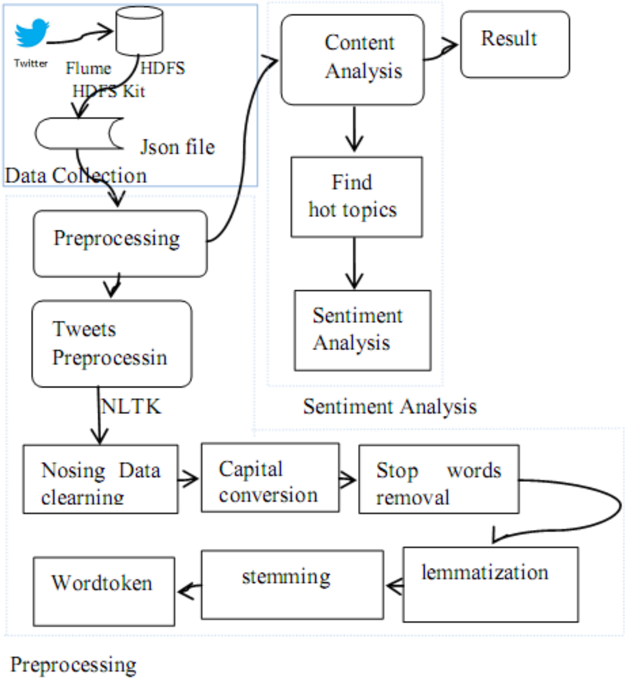


Fig. 1. The work flow.

Data normalization: Preprocess the collected tweets data through 6 sub-steps using natural language processing. Which is nosing data cleaning, Capital conversion, Stop words removal, lemmatization, stemming and word token.

Sentiment analysis: According to the frequency of words, we picked five hot topics about HUAWEI smartphones. Finally, chose one to analyze.

Conclusions: Summarize the result and discuss the influence according to our analysis.

A. Data Collection

In this paper, using Apache Flume to collect tweets data from Twitter. Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into Hadoop Distributed File System (HDFS). Using Flume to collect data steps as shown in Fig. 2.

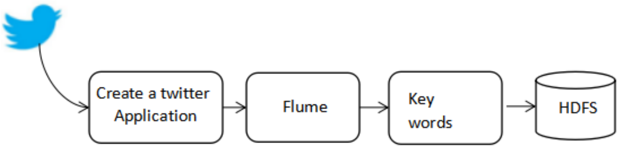


Fig. 2. Collect data from Twitter using Flume.

Firstly, created a Twitter application, and get consumer key, consumer secret, access token, access secret. And then configuration the Flume configuration file. The configuration

Manuscript received July 6, 2018; revised August 19, 2018.

Xiaorui Shao is with Information collaboration Engineering department, Pukyong national university, Korea (e-mail: sxrpknu@qq.com).

Chang-Soo Kim is with the IT convergence and application engineering Department of Pukyong national university, Korea (e-mail:cskim@pknu.ac.kr).

Kwak Dong Ryul is with PKNU Department of IT Convergence and Application Engineering, Korea (e-mail:destroy247@naver.com).

DOI: 10.7763/IJCTE.2019.V11.1232

file as shown in the Table I.We use 'HUAWEI smartphone' as a keyword to collect tweets. It created one folder daily and stored in HDFS. And when the file size more than 200MB, it will create a new file. We use Flume to collect 1.4GB data from twitter .192390 tweets.

TABLE I: FLUME CONFIGURATION FILE

```
TwitterAgent.sources=Twitter
TwitterAgne.channels=MemChannel
TwitterAgent.sinks=HDFS
TwitterAgent.sources.Twitter.type=org.apache.flume.twitter.TwitterSources
TwitterAgent.sources.Twitter.consumerKey=YOUR CONSUMERKEY
TwitterAgent.sources.Twitter.consumerSecret=YOUR CONSUMERSECRET
TwitterAgent.sources.Twitter.accessToken=YOUR ACCESSTOKEN
TwitterAgent.sources.Twitter.accessTokenSecret=YOUR ACCESSTOKEN SECRET
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=YOUR PATH(host:8820/your path)
TwitterAgent.sinks.HDFS.hdfs.filePrefix=log_%Y%m%d_%H
TwitterAgent.sinks.HDFS.hdfs.useLocalTimeStamp=true
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=100
TwitterAgent.sinks.HDFS.hdfs.rollSize=204800000
TwitterAgent.sinks.HDFS.hdfs.rollCount=0
TwitterAgent.sinks.HDFS.hdfs.rollInterval=0
TwitterAgent.sinks.HDFS.hdfs.minBlockReplicas=1
TwitterAgent.ChannelsMemChannel.type=memory
TwitterAgent.ChannelsMemChannel.capacity=10000
TwitterAgent.ChannelsMemChannel.transactionCapacity=1000
TwitterAgent.sources.Twitter.channels=MemChannel
TwitterAgent.sinks.Twitter.channels=MemChannel
```

B. *Data Normalization*

Twitter data are very noisy, unstructured and complicated [4]. This paper mainly about contents analysis of tweets. The tweets contain symbols, link, emotions, etc. To extract useful information to analyze, preprocessing [5] tasks are very important and critical in text mining [6]. We adopt five steps to normalize the tweets text using natural language processing. And we set an example to explain these steps. As shown in Fig. 3.

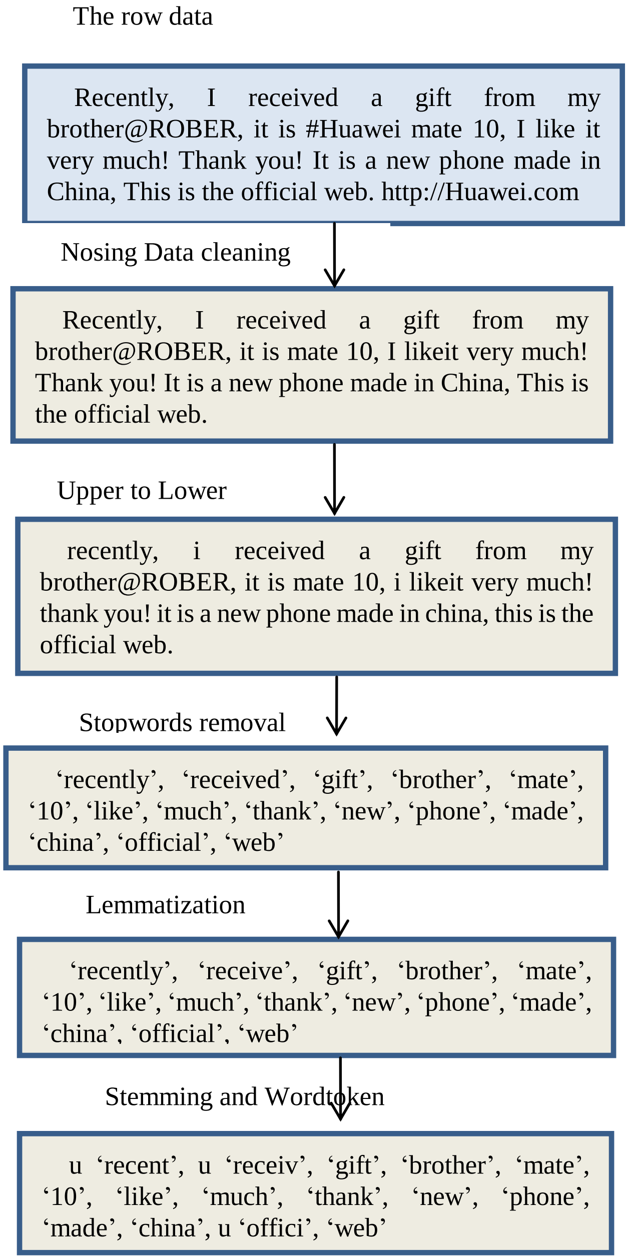
Nosing data cleaning: Collected tweets contain a lot of noisy information, Like username@ hash tags#, and URL. In this paper, we remove the username@, hashtags and URL firstly.

Upper to lower: When we analyze the text, the tweets text contain some upper letter. Converting all of the letters into lower is necessary.

Stop words removal: In tweets, a lot of stop words are contained in the content. Such as 'the,' 'are,' 'in,' 'one,' and 'is' etc. They are not important and useless for sentiment analysis.

Lemmatization: In the English language, most of the words having many types, such as 'make,' 'made,' 'makes.' We need to extract the normal type among several types.

Stemming and word token [7]: Like lemmatization, to reduce inflectional forms in tweets, like change 'recently' into 'recent,' 'mainly' into 'main.' In this paper, we choose Porter stemmer [8] to stem.



The row data

Recently, I received a gift from my brother@ROBER, it is #Huawei mate 10, I like it very much! Thank you! It is a new phone made in China, This is the official web. http://Huawei.com

Nosing Data cleaning

Recently, I received a gift from my brother@ROBER, it is mate 10, I likeit very much! Thank you! It is a new phone made in China, This is the official web.

Upper to Lower

recently, i received a gift from my brother@ROBER, it is mate 10, i likeit very much! thank you! it is a new phone made in china, this is the official web.

Stopwords removal

'recently', 'received', 'gift', 'brother', 'mate', '10', 'like', 'much', 'thank', 'new', 'phone', 'made', 'china', 'official', 'web'

Lemmatization

'recently', 'receive', 'gift', 'brother', 'mate', '10', 'like', 'much', 'thank', 'new', 'phone', 'made', 'china', 'official', 'web'

Stemming and Wordtoken

u 'recent', u 'receiv', 'gift', 'brother', 'mate', '10', 'like', 'much', 'thank', 'new', 'phone', 'made', 'china', u 'offici', 'web'

Fig. 3. Data normalization.

III. SENTIMENT ANALYSIS

We had collected data and preprocessed them. There we only analyze the English tweets, and we must acknowledge not all English tweets made by customers. Also, the HUAWEI Official Twitter can push tweets, but compared with so many users, it is just a small part. So, the analysis is also effective. Firstly, we picked ten hot topics among these tweets. And then according to these hot topics, we can do further analysis.

A. *Topic Naming*

Following the steps defined in subsection Ⅱ.*B*, we obtained the following result depicted as a word cloud as shown in Fig. 4. We set the max number of words shown in word cloud is 50.

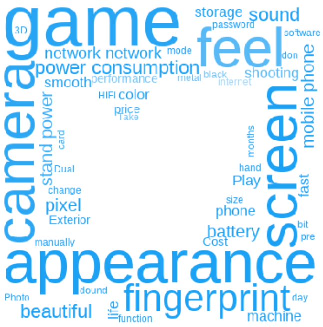


Fig. 4. The obtained main topic.

However, we find there is still having some meaningless words, such as 'feel,' 'play,' 'take.' At tweets text normalization steps, we have removed the stop words, but there we need to create another stopwords dictionary. According to our analysis demand, we use a stopwords dictionary which consists of 908 words. The final results as shown in Fig. 5. For more current and clear, we set the max number of the word shown in word cloud is 10.

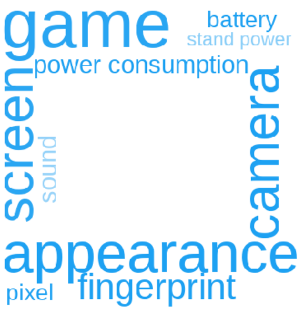


Fig. 5. The processed main topic.

Through this picture, we can know the customers' primary concern is game, the phone's appearance, camera, fingerprint, screen, standby power consumption, battery, sound, and pixel.

B. *Customers' Sentiment Analysis*

According to previous results, we choose HUAWEI honor 9 (a kind of HUAWEI smartphone) to analyze, all of the tweets is 16636. First, divide customs' emotions into three categories: positive, negative, neutral. The Table II gives an example. First, we label 2000 tweets as positive, negative, neutral. And then randomly divided them into training sets and test sets. Specifically 30% test sets, 70% training sets. According to the accuracy of three classifiers: Decision Tree (DT) Classifier, Naive Bayes (NB) Classifier,

Maximum Entropy (ME) Classifier [9]. As shown in Table III, we Choose Decision Tree Classifier to classify.

TABLE II: THE EMOTION LABEL BY HANDING

Tweets sample	Emotions
HUAWEI honor smartphone has a good price.	Positive
How to recover lost data from Huawei honor?	Neutral
Camera 13mp 2mp ram 4gb battery capacity HUAWEI honor 9 lite full specifications and price is so low.	Positive
I don't like this phone.	negative

TABLE III: THE ACCURACY OF THE THREE CLASSIFIERS

Classifier	Accuracy
Decision Tree	0.7883
Naive Bayes	0.7063
Maximum Entropy	0.5813

Firstly, analyze customers overall emotion for HUAWEI honor nine smartphones. The result as shown in Fig. 6. We know most people are satisfied with it. But the negative emotion is still taking up 36.2%.

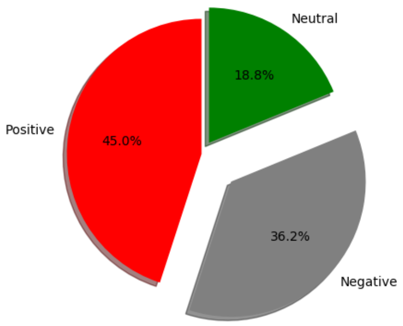


Fig. 6. Customers' emotion.

Next, we analyze the customers' emotion for phone's pixel, sound, storage, game, standby power consumption, screen, game, network, resolution, and battery. The result as shown in Fig. 7.

We can get a positive evaluation about the pixel, storage, and resolution over 55%. It represents most of the customers are satisfied with it. And about sound, screen, network, the ratio of positive and negative evaluation almost is similar, the customers' emotion depends on themselves. But for the game, standby power consumption and battery, obviously, over 60% of customers give a bad evaluation. There is needed to improve a lot.

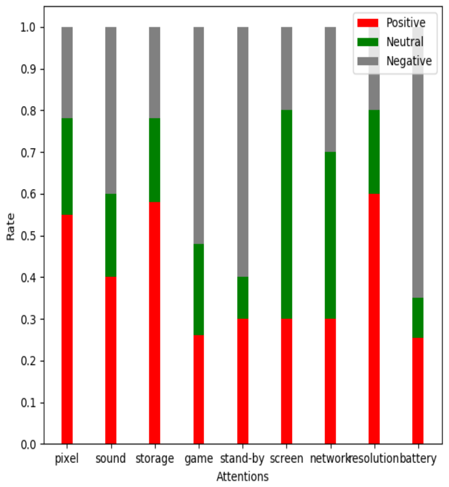


Fig. 7. Positive, negative, neutral ratio situation.

IV. RESULTS

In this paper, we described the whole procession about how to analyze the customers' emotional behaviors using Twitter data. Including data collection, data preprocessing, and sentiment analysis. Through analyze HUAWEI smartphone. We know how to choose the best classifier to classify the tweets text. The main emotion is positive; it takes up to 45%. However, negative is still taking up 36.2%. For special analysis, in the field of battery, standby power consumption, and game, almost over 60% of the customer are not satisfied with it. HUAWEI Company must make huge progress.

REFERENCES

[1] A. Biffet and E. Frank, "Sentiment knowledge discovery in Twitter streaming data. Discovery science," *Comput Science*, vol. 6332, pp. 1-15, 2010.

[2] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford University Project Technical Report, 2009.

[3] E. Kalampokis, "Understanding the predictive power of social media," *Internet Research*, vol. 23, pp. 544-559, 2013.

[4] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing tasks in Indonesian twitter messages," *Journal of Physics: Conference Series*, 2018.

[5] Mastering social media mining with python. (2016). *Marco Bonzanini: Python, Data Science, Text Analytics*. [Online]. Available: https://marcobonzanini.com/2015/03/09/mining-twitter-data-with-python-part-2/

[6] D. Torunoğlu, E. Çakırman, M. C. Ganiz, S. Akyokuş, and M. Z. Gürbüz, "Analysis of pre-processing methods on classification of Turkish texts innovations in intelligent systems and applications (INISTA)," in *Proc. the 2011 International Symposium on IEEE*, 2011.

[7] Stemming and lemmatization. (2013). [Online]. Available: http://blog.csdn.net/march_on/article/details/8935462

[8] C. J. van Rijsbergen, S. E. Robertson, and M. F. Porter, "New models in probabilistic information retrieval," London: British Library Research and Development Report No. 5587.

[9] API reference — Blob classes. [Online]. Available: https://textblob.readthedocs.io/en/dev/api_reference.html#api-classifier

Xiaorui Shao is a M.S. candidate in the Information Collaboration Engineering department, Pukyong National University, Korea. His current research interests include big data, smart factory, and machine learning.

Chang Soo Kim is a full professor of IT Convergence and Application Engineering, Pukyong National Universty, Korea. His skills and expertise include information communication technology, information technology, cloud computing, data mining and knowledge discovery, information system management, IT project management, social networking, social network analysis, information extraction, computer science education, web science, web mining, IT governance, information technologies, social media, information extraction, data warehousing, information systems engineering, service oriented architecture and so on. He has more than 50 high-quality papers. He got his PHD degree in Sung Kyun Kwan University.

Kwak Dong Ryul is a student from PKNU Department of IT Convergence and Application Engineering. His current research interests include big data and machine learning.