IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. III (Sep. - Oct. 2016), PP 54-60
www.iosrjournals.org

Opinion Mining Method for Sentiment Analysis

Sristi Sharma[1], Dr. Surendra Kumar Yadav[2], Mr. Lokendra Pal[3]
[1](Computer Science and Engineering/ JECRC University, India)
[2](Computer Science and Engineering/ JECRC University, India)
[3](Computer Science and Engineering/ JECRC University, India)

Abstract: *We are living in a world full of data. Every passing second, large data is generated by Social Media, E-Commerce, Stock Exchange and many other platforms. Now-a-days, microblogging sites are used for manypurposes, such as communication, trend detection in marketing and business products, sentiment analysis, election prediction, education and much more, which has changed the public perspective of personalization and socialization. Twitter, is one of the major sources of data which generates more than 500 million of tweets per day. Every user usually shows his feeling or emotion about the topics of his interest. The reason of information gathering is to find out what the people feel. The decurtate length and the highly colloquial nature of tweets render it very difficult to automatically detect the sentiment and thus the requirement of sentiment analysis. Sentiment Analysis and summarization has in recent years caught the attention of many researchers, as on line text analysis is highly beneficial and asked for in various applications. A typical application is product-based sentiment summarization. This multi-document summarizationinforms users about pros and cons of available products. Sentiment analysis allows box office, social media, business analytics, market and FOREX rate prediction, and is also used in recommender system. In the present era sentiment analysis is most interesting research topic in text mining in the field of NLP in which a valuable knowledge extraction from textual data posted on the social media is an onerous task. A new framework has been proposed in this paper to normalize the text and judge the polarity of textual data aspositive, negative or neutral using an ETL (Extract, Transform, and Load) big data tool called Talend. The algorithm developed focusses on parallelism for performance speed and contributes towards the end result by comparing the accuracy with standard data set .*

Keywords: *Talend, sentiment analysis, lexicon, AFFIN, sentiwordnet.*

I. Introduction

In the Internet era, it is much easier to collect diverse opinions from different people around the world. People look to review sites (e.g., CNET, Epinions.com),e-commerce sites (e.g., Amazon, eBay), online opinion sites (e.g., TripAdvisor, Rotten Tomatoes, Yelp) and social media (e.g., Facebook, Twitter) to get feedback on how a particular product or service may be perceived in the market. Similarly, organizations use surveys, opinion polls, and social media as a mechanism to obtain feedback on their goods and facilities.

Sentiment analysis and opinion mining are area of study that helps in analysis of people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the highly pursued research areas in natural language processing. It is also widely studied and used in data, Web, and text mining. Today it is a widely accepted research area not only in computer science but also in management sciences and social sciences due to its importance to business and society as a whole. Sentiment analysis has grown and important techniques coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks.[1]

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state or the intended emotional state. Our main focus is on to find the opinion mining method or framework to analyze the sentiments of twitter.[2]

II. Literature Survey

Information exchange using social networking like facebook , blogs, twitters etc. are an important part of human life. However, today we have graduated much beyond information exchange and we now analyze the available data using various techniques. One such technique is sentiment analysis which is field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. Sentiment Analysis helps in achieving various goals like observing public mood regarding political movement, market

intelligence, the measurement of customer satisfaction, movie sales prediction and many more. Number ofpapers are available on the subject. A survey has been carried out and their review is given in the succeeding paragraphs.[3]

A number of approaches have been applied in sentimental analysis for increasing the accuracy of result. Turney, 2002 [6] used an unsupervised learning algorithm for classifying the algorithm . He then categorized 410 reviews into automobiles ,banks, movies, travel destinations domains and after the algorithm achieved an accuracy of 84%,80%,68.83% and 70.53% respectively. The overall accuracy of all the domains achieved was 74.39%. (Pang et al.,2002) [7] developed an algorithm and tested the same on the reviews of movies. He used maximum entropy classification, Naive Bayes and support vector machines and achieved the accuracy of 80.4%, 81% and 82.9% respectively. He observed that support vector machine has the best results as compared to other machine learning algorithms.(Alec Go et al., 2009) [9] carried out sentiment analysis on twitter dataset of 16 lac tweets using Naive Bayes, and support vector machines learning algorithms. he achieved an accuracy of 80%.

(Bifet and Frank, 2010) [8] used the sliding window Kappa statistics on twitter time changing data streams. He used different machine learning algorithms namely Stochastic Gradient Descent, Multinomial Naive Bayes, and Hoeffding Tree on two different data stream statistics to analyze the result. . (Agarwal et al., 2011) [10] examine the sentiment analysis using the POS specific prior polarity features and uses two new pre-processing resources i.e. emoticon dictionary and acronym dictionary on 11,875 manually annotated Twitter data from a commercial source and got the accuracy of 73.4% with F-measure for positive and negative sentiment is 71.13% and 71.50% respectively on unigrams.(Mudinas et al., 2012) [11] carried out experiment on reviews of software and movies using a pSenti (hybrid) approach which an the combination of lexicon-based and learning-based approach and achieved the accuracy more than 78% in software reviews and 82.3% in movies reviews approaches. (Soo-Guan Khoo et al., 2012) [12] used the appraisal theory on political news article and try to identify different aspects of sentiments such as attitude ,emotions and bias of the appraisals of the author.

(Sunil B. Mane et al., 2014) [13] performed sentiment analysis on 1466 tweets using Hadoop and obtained 68.40% accuracy. (Anthony M. Hopper and Maria Uriyo, 2015) [14] applied the Time-to-next-complaint analytical methods to review patient comments feedback to unable hospital managers to view sentiments and derive valuable information to satisfy the queries of patients.

III. Proposed Approach

The proposed approach is a dictionary based technique i.e. a dictionary of sentiment bearing words along with their polarities has been used to classify the text into positive, negative or neutral opinion. We have applied the method of sentiment scoring using standard dictionaries like the SentiWordNet ,Opinion Lexicon word list ,AFINN polarity word list and assigned the polarity value for the matched tokens (words) and further by which we can reduce the neutrality. The steps used as shown below.

- Step-1 – Load the dataset of the Twitter having 1466 tweets(Text file).
- Step-2 – Pre-processing is done to remove Stop-words, Hash-tags, @user name, Re-Tweet, Hyper-link, Punctuation symbol and expand smiley's, Acronym, use of NER Tagger[21].
- Step-3 – Sentiment classification is done by using SentiWordNet [17] into positive, negative or neutral sentiments.
- Step 4 – If neutral sentiments occur then re-classification takes place by breaking into tokens and then apply is Opinion Lexicon dictionary[18][19] and AFFIN polarity dictionary[20] to determine the polarity.
- Step 5 – Result are shown in tabular and visual form.

3.1 Flow Diagram

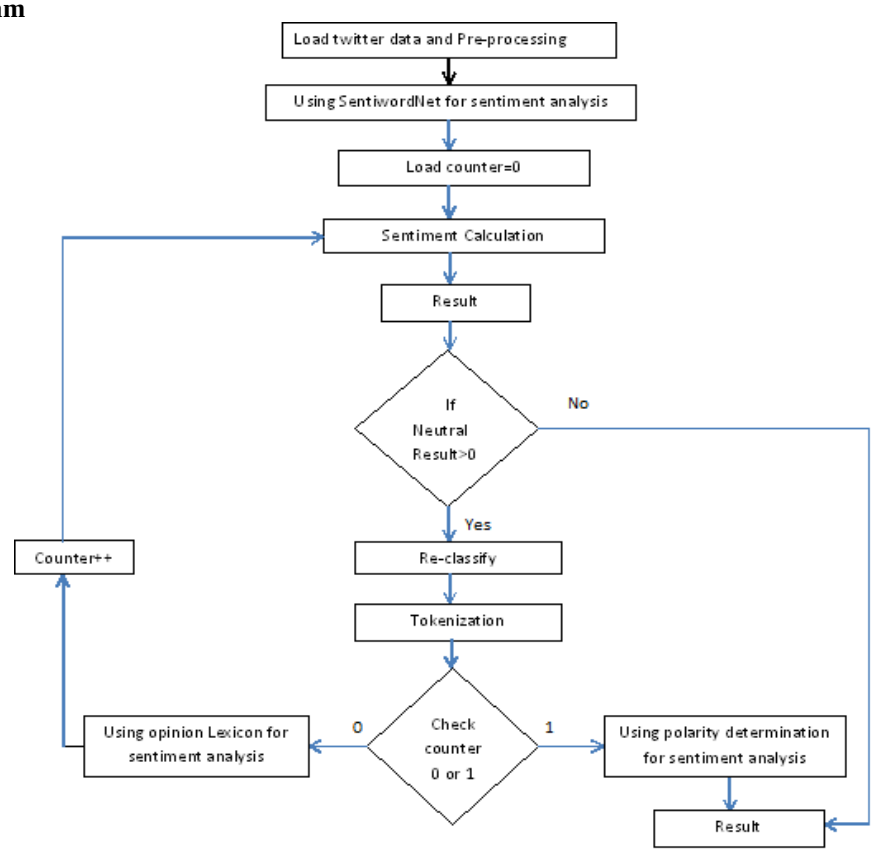

Figure 3.1 : Flow Chart of Proposed Work

IV. Proposed Algorithm

Let

T : tweet
TL : Data List of tweet T
PP : Pre-process function to detect smiley's, to remove Stop-words, Hash-tags, @user name, Re-Tweet, Hyper-link, Punctuation symbol and digits Acronym slang detection, use of NER Tagger
PPT : Pre-processed tweet T
SWN : SentiWordNet Function to compute sentiment weight for each tweet.
PPTi : ith word of Tweet PPT
S(PPT) : Sentiment weightage of Tweet PPT
S(PPTi): Sentiment weightage for ith word of tweet.
PWL : Data List of Positive words in Opinion lexicon word list
NWL : Data List of negitive word in Opinion lexicon word list
APWL : AFINN polarity words List.
APWS : AFINN polarity words weightage List (pre-defined weightage at i'th index for i'th word in APWL).

Let posCount = 0;
Let negCount = 0;
Let neuCount = 0;
for tweet T in TL
PPT = PP(T) **//pre-Processed tweet using function PP**
S(PPT) = SWN(PPT) **//weightage calculated using SWN function**

```
/* If tweet is neutral , re-classify using Opinion Lexicon*/
        if(S(PPT)==0) then
                foreach word PPTi in PPT do
                                /*Checking positive lexicon*/
                        if(PPTi in PWL) then
                                S(PPT) = S(PPT)++
                        end if
                                /*Checking negative lexicon*/
                        if(PPTi in NWL) then
                                S(PPT) = S(PPT)--
                        end if
                end for
        end if

/* For neutral tweets after using Opinion Lexicon, re-classify using AFFIN Polarity Dictionary*/
        if(S(PPT) == 0) then
                foreach word PPTi in PPT do
                        if(PPTi in APWL) then
                                S(PPTi)= APWS.get(APWL get index of PPTi)
                                S(PPT)=S(PPT)+S(PPTi)
                        end if
                end for
        end

        if(S(PPT) > 0)
                posCount++
        else if(S(PPT) < 0)
                negCount++
        else
                neuCount++;
        end else if
end for
```

V. Performance Evaluation

In this section, description of the tool which has been used to perform sentiment analysis is given along with the performance of the proposed approach and its comparison with existing works.

5.1experimental setup

The proposed algorithm has been implemented using big data tool Talend 6.6.1[22]. Talend [5]was founded by BertrandDiard and Fabrice Bonan in 2005. It is open source software that provides big data, Cloud, Master Data Management, data quality, data integration, data management and enterprise application integration software and services. It has 4500+ connectors and components which allow multiple input/outputs in a single job. It is an ETL tool. ETL process describes the three components: Extraction, Transformation, Loading.

5.2 dataset used

Experiment have been carried out on an entries of twitter dataset. It have five fields i.e. User ID , tweet ID, name, sentiment, tweet and 1466 entries[13]. This standard dataset after processing is having 732 positive polarity tweets ,730 negative polarity tweets and 4 neutral entries. The dataset is available in the link i.e**http://www.cs.tau.ac.il/~kfirbar/mlproject/twitter.data**

5.3 dictionary used

We have applied the method of sentiment scoring using the dictionaries viz. SentiWordNet[17] ,Opinion Lexicon word list[18],[19] ,AFINN polarity word list[20] and have assigned the polarity value for the matched tokens (words) by which we can reduce the neutrality.

5.4 WorkFlow

We have applied the method of sentiment scoring using the SentiWordNet[17] for classifying the tweet as either positive (+1), negative (-1) or neutral (0) ,the number of neutrals is high then we re-classify by breaking the tweets into tokens (words) and match then with the Opinion Lexicon word list [18],[19] and assign the score +1 if matched with a positive word list, -1 if matched with a negative word list and 0 if doesn't match with any word list. Thereafter the score is classified as either positive (+1), negative (-1) or neutral (0) by adding the sentiment score of each token group by their respective id, and in this process if their score is greater than zero then it is positive, less than zero then it is negative or neutral if it is equal to zero. If still the score of the neutral is high then all the neutral tweets are re-classified again by breaking the tweets into tokens (words) and matched with the AFINN polarity word list [20] and assign the polarity value for the matched tokens (words) by which we can reduce the neutrality.

VI. Result

6.1 accuracy analysis of our proposed algorithm

In our experiment we have used the standard dataset[13] having three sentiments positive , negative and neutral , and having 732 count,730 count, and 4 count respectively. After applying our developed experimental algorithm on the dataset the following results have been obtained i.e. 772 positive, 595 negative and 94 neutral. This has been compared with the standard dataset result and the result of the of the positive count is 629, negative count is 539 and neutral count is 1. With help of above values we have calculated the accuracy of the sentiments as 85.93%, 73.84 % and 25% respectively. In our experiment to calculate the total accuracy we map total dataset count and our obtained count from the experiment and get accuracy of 79.74 %. The same is shown in the table 6.1

Sentiment	Dataset count	Result of proposed algorithm	Accuracy
Positive	732	629	85.93%
Negative	730	539	73.84%
Neutral	4	1	25%
Total	1466	1169	79.74%

Table 6.1: Result of proposed algorithm

To test our experimental accuracy we have compared our result with the Result of Mane et al.[13] and Result of P.Alexander et al[23](reference papers). The result of Mane et al.[13] is having 1004 experimental count as against a dataset of 1466 counts and Result of P.Alexander[23] is having 216 experimental count as against a dataset of 375 counts. Accuracy has been calculated by mapping the above values with the standard dataset. An accuracy of 68.40% is obtained for Mane et al. and 57.6% is for P.Alexander et al. The same are shown at table 6.2

Sentiment	Dataset count	Result of Mane et al.	Accuracy
Positive	729	542	74.04%
Negative	665	458	68.87%
Neutral	72	53	73.61%
Total	1466	1004	68.40%
Sentiment	**Dataset count**	**Result of P.Alexander et al.**	**Accuracy**
Positive	160	108	67.5%
Negative	125	75	60%
Neutral	90	33	36.66%
Total	375	216	57.6%

Table6.2 : Result of Mane et al. and P.Alexander et al.

Comparative assessment of sentiments count of our experiment and reference papers is at table 6.3. The graphical representation of the same is at figure 6.1

	Positive	Negative	Neutral
Result of proposed algorithm	85.93%	73.84%	25%
Result of Mane et.al	74.04%	68.87%	73.61%
Result of P.Alexander et.al	54.54%	50%	31.42%

Table6.3 : Calculation of Accuracy according to category

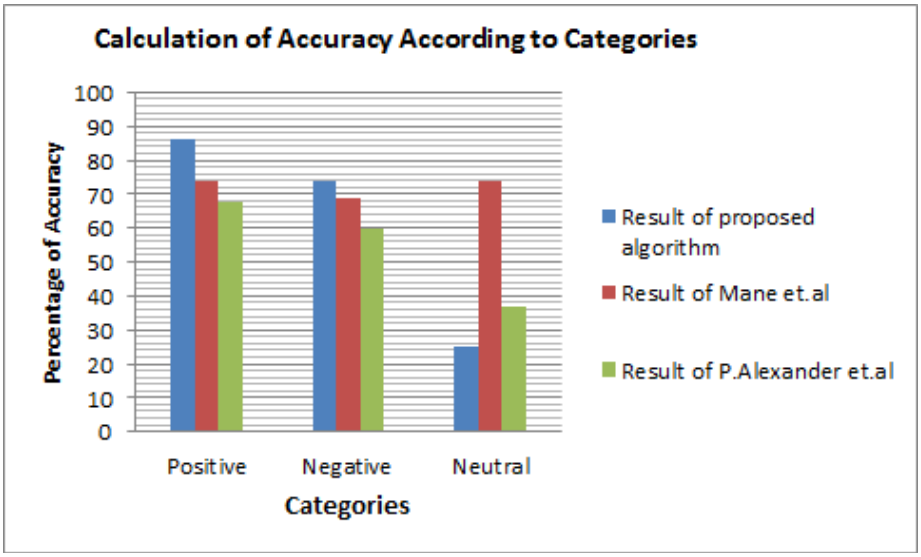

Figure6.1 : Calculation of Accuracy according to category

6.2 Comparison of overall Accuracy of Algorithm

The comparative analysis of the overall accuracy of the algorithm with the referencepapers is shown in the table 6.4 the same is graphically represented in figure 6.2

	Result of proposed algorithm	**Result of Mane et al.**	**Result of P.Alexander et al.**
Comparison of overall Accuracy of algorithm	79.74	68.40	57.6

Table 6.4: Comparison of overall Accuracy of algorithm

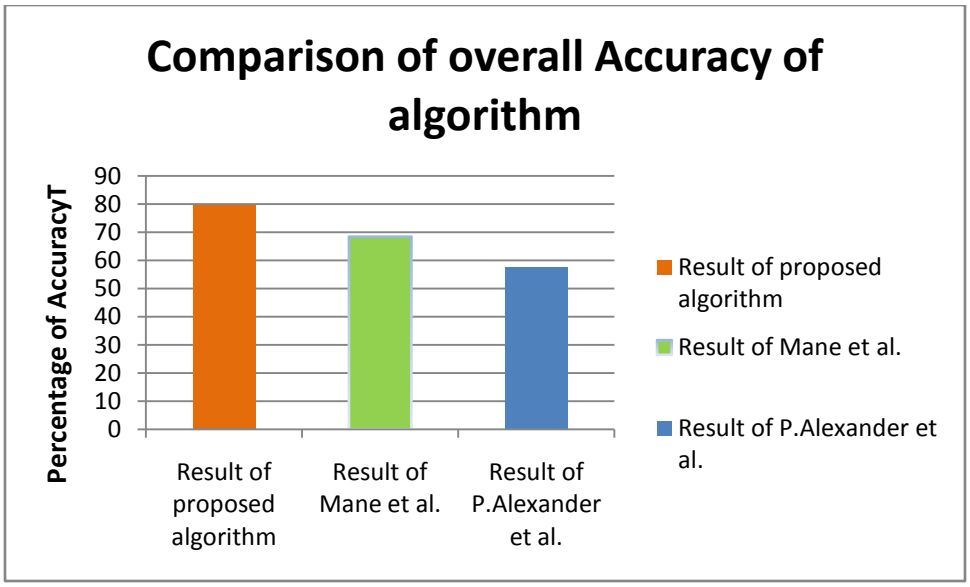

Figure6.2 : Comparison of overall Accuracy of algorithm

VII. Conclusion And Future Work

Sentiment analysis and opinion mining are area of study that helps in analysis of people's opinions, sentiments, evaluations, attitudes, and emotions from written language. A number of algorithms have been developed and results obtained have been published in papers as enumerated in references. A sincere effort has been made to develop an algorithm to carry out similar analysis and thereafter a comparative assessment has been carried out with non-research algorithms. The main focus of our work is to improve accuracy which has been validated by

the test results. But the work can be expanded by introducing techniques that further increase the accuracy. The developed algorithm takes 0.99 sec per tweet to execute which is considered more.Thiscan also be shortened by establishing some time improvement technique.

References

[1]. https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdfBing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.BLiu -Synthesis lectures on human language technologies, 2012 - morganclaypool.com

[2]. About sentimental analysis is available on the link- https://en.wikipedia.org/wiki/Sentiment_analysis

[3]. Ravi, Kumar, and Vadlamani Ravi. "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems 89 (2015): 14-46.

[4]. Talend: About Talend. https://www.talend.com/about-us.

[5]. Turney, Peter D.Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417-424. Association for Computational Linguistics, 2002.

[6]. Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002.

[7]. Bifet, Albert, and Eibe Frank. "Sentiment knowledge discovery in twitter streaming data." In Discovery Science, pp. 1-15. Springer Berlin Heidelberg, 2010.

[8]. Go, Alec, RichaBhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.

[9]. Agarwal, Apoorv, BoyiXie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In Proceedings of the workshop on languages in social media, pp. 30-38. Association for Computational Linguistics, 2011.

[10]. Mudinas, Andrius, Dell Zhang, and Mark Levene. "Combining lexicon and learning based approaches for concept-level sentiment analysis." In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, p. 5. ACM, 2012.

[11]. Soo-Guan Khoo, Christopher, ArminehNourbakhsh, and Jin-CheonNa."Sentiment analysis of online news text: a case study of appraisal theory." Online Information Review 36, no. 6 (2012): 858-878.

[12]. Mane, Sunil B., YashwantSawant, SaifKazi, and VaibhavShinde. "Real Time Sentiment Analysis of Twitter Data Using Hadoop." Int. J. Comput. Sci. Inf. Technol 5, no. 3 (2014): 3098-3100.

[13]. Hopper, Anthony M., and Maria Uriyo. "Using sentiment analysis to review patient satisfaction data located on the internet." Journal of health organization and management 29, no. 2 (2015): 221-233. Harvard.

[14]. Hridoy, Syed Akib Anwar, M. TahmidEkram, Mohammad Samiul Islam, Faysal Ahmed, and Rashedur M. Rahman. "Localized twitter opinion mining using sentiment analysis." Decision Analytics 2, no. 1 (2015): 1-19.

[15]. Twitter Application Management. https://apps.twitter.com/.

[16]. Baccianella, Stefano, Andrea Esuli, and FabrizioSebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In LREC, vol. 10, pp.2200-2204. 2010.

[17]. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177. ACM, 2004.

[18]. Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." In Proceedings of the 14th international conference on World Wide Web, pp. 342-351. ACM, 2005.

[19]. Hansen, Lars Kai, Adam Arvidsson, Finn rup Nielsen, Elanor Colleoni, and Michael Etter. "Good friends, bad news-affect and virality in twitter." In Future information technology, pp. 34-43. Springer Berlin Heidelberg, 2011.

[20]. Finkel, Jenny Rose, TrondGrenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling." In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363-370. Association for Computational Linguistics, 2005.

[21]. TALEND DOWNLOADS "Download Talend Open Studio" Available: https://www.talend.com/download/talend-open-studio.

[22]. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.