Determining the mood of social network users in real-time by sentiment analysis methods using Big Data tools

Rostyslav Ivankiv, Ihor Lazarovych

Vasyl Stefanyk Precarpathian National University
Ivano-Frankivsk, Ukraine

I. Introduction

As the number of Internet users increases, so does the amount of content they generate. All this content carries a large amount of data that can and should be used for further analysis. To do this, there is a separate area of artificial intelligence - natural language processing, and one of its promising areas is sentiment analysis.

The purpose of the work is to determine the mood of users about a particular topic in real-time. It is necessary to find the optimal method of determining the mood of social network users, using approaches to sentiment analysis in the context of working with large amounts of data.

Firstly, a common problem with big data analytics is that traditional analysis systems are not reliable to process large amounts of data at an acceptable rate. Secondly, big data processing usually requires costly processes of cleaning, pre-processing and data conversion, as data is available in different formats, both semi-structured and unstructured. Finally, big data is constantly generated at high speed, which means that none of the traditional data preprocessing architectures is suitable for efficient real-time analysis. The paper describes general practices for the current subject and proposes improvements for sentiment analysis methods in the real-time streaming conditions and further processing by means that allow you to scale on the total load.

II. Analysis of recent research and publications

There are a lot of different research for the current topic which covers some specific designed solution. Essentially they can be classified as follows: sentiment analysis research by itself (Lexicon analysis, Machine learning [1], Deep Learning [2]) and sentiment analysis research in terms of Big Data (Batch processing, Near Real-time processing, Real-time processing).

For example, the next publication [3] described the various sentiment analysis techniques and their general methodology using MapReduce on a Hadoop, which is performance limited because of historical reason while selecting storage architecture on disk only. In most of the works, an analysis was performed on static data, but in the next work [4] discovered near real-time processing improvements. The other authors [5] created a finding patterns methodology related to the health events where they tried to analyze it using the approach upon classifying the list of words. Such authors as Nguyen and Jung [6] was offering a method of event detection through the behavioral analysis of Twitter users by utilizing real-time data analytics on big social data. Also, an architecture [7] was developed for social media text analysis by keywords filtering, languages, and other informative aspects of a large number of tweets.

III. Presetation of the material

Determining the mood of social network users can be done in the way of using techniques for sentiment analysis after you have relevant and clean data.

The tonality of the text is determined by the lexical tonality of its constituent units and rules for their combination. A tonal score can be presented in one of the following types: binary (positive / negative), ternary (positive / neutral / negative), ranged.

It is proposed to search for lexical sentiment in the text according to pre-compiled tonality dictionaries (lists of patterns). In the dictionary, each word corresponds to tonal assessment. Such an indicator is a set of five values (V_1, V_2, V_3, V_4, V_5). Each word determines the degree of belonging of the word to one of the classes: very negative(V_1), negative(V_2), neutral(V_3), positive(V_4), very positive(V_5). The sum of all values for a particular word is equal to 1. In the case if word is absent in the dictionary, its tonality is considered neutral. For the final tone assessment of the whole text, is needed to calculate the ratio of these components using the formula [8]:

$$S = \frac{\sum_{i=1}^{N} V4_i + V5_i}{\sum_{i=1}^{N} V1_i + V2_i} , \qquad (1)$$

where S - tone score of the sentence; $V = [V_1, V_2, V_3, V_4, V_5]$ - tone score of the word; V_1, V_2 - negative components; V_4, V_5 - positive components; V_3 – neutral component; N - number of words in the sentence.

Based on the results of the experiments the accuracy of sentiment determination is calculated as follows [9]:

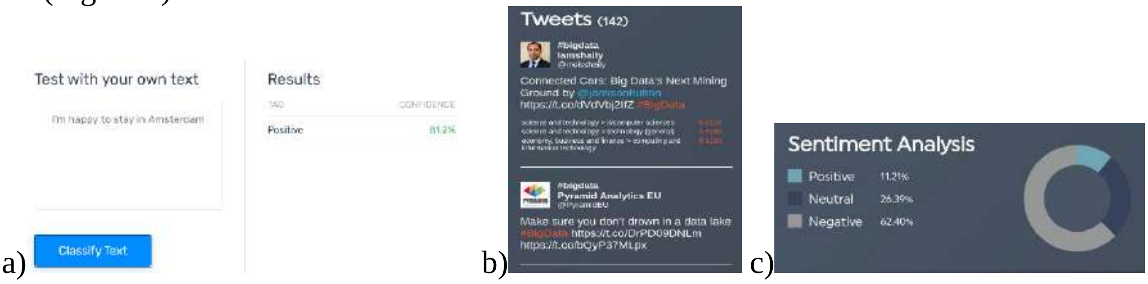
$$A = \frac{N_c}{N_a} , \qquad (2)$$

where A – accuracy of the sentiment determination; N_c - number of experiments with correctly defined sentence tones; N_a - total number of experiments.

Table 1 – Accuracy evaluation results

Dataset	Accuracy
Twitter dataset with 1.6 million tweets[10]	0.73
Historical Twitter Datasets[11]	0.70
Facebook comments dataset[12]	0.65
Twitter US Airline Sentiment dataset[13]	0.81

The described approach to analyze tone of a text is possible to use and integrate into the application (Figure 1).



Test with your own text

I'm happy to stay in Amsterdam

Results

TAG	CONFIDENCE
Positive	81.2%

Tweets (142)

#bigdata
Iamshaily

Connected Cars: Big Data's Next Mining Ground by @jamisonhutton https://t.co/dVdVbj2IfZ #BigData

#bigdata
Pyramid Analytics EU
@PyramidEU

Make sure you don't drown in a data lake #BigData https://t.co/DrPD09DNLm https://t.co/bQyP37MLpx

Sentiment Analysis

Positive	11.21%
Neutral	26.39%
Negative	62.40%

Figure 1. Sentiment analysis app view

The creation of a basic pipeline (Figure 2) is required for a better understanding of the general view of the components and their potential usage. Data was streamed using Twitter API and other social network resources, real-time data was pushed and assembled in Kafka messaging queue for further processing on Spark - analytics engine for large-scale data processing. Once it has been processed there is an opportunity to visualize the results using different frameworks. By utilizing such Big Data tools as Kafka and Spark it is possible to scale horizontally accordingly to needs which allows working on a high throughput with high efficiency. Although sometimes it also not a good fit. For instance, when you have low computing capacity since Spark uses cluster memory, you should go for other alternatives, such as Apache Hadoop which uses a disk instead. Also Kafka stores redundant copies of data, which can increase storage costs. Kafka is originally designed to cope with the high load so that usage is an overkill when you need to process only a small amount of messages per day (up to several thousand). Recommended using traditional message queues like RabbitMQ in such cases.

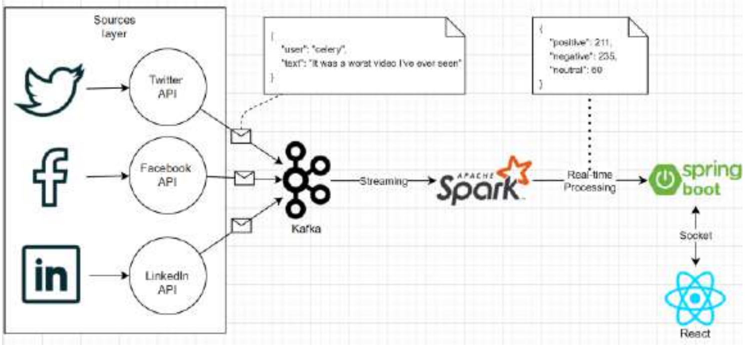

Sources layer
Twitter API
Facebook API
LinkedIn API
{ "user": "celery", "text": "It was a worst video I've ever seen" }
{ "positive": 211, "negative": 235, "neutral": 69 }
Kafka
Streaming
Apache Spark
Real-time Processing
spring boot
Socket
React

Figure 2. Real-time sentiment analysis pipeline

IV. Conclusions

Using dictionaries that were used in current work most likely is the simplest possible way to perform sentiment analysis. Even though sometimes it is possible to use for many cases, it still fails to handle all the complexities of a language. Under circumstances that prediction accuracy is critical and if training data is available, it does make sense to use the ML-based approach which often outperforms dictionary-based methods. In general, the following approach for real-time processing which requires a bunch of data and resources to handle is considered to be valid and helpful with the usage such well-designed, high-throughput, fault-tolerant facilities as Kafka and Spark, since that makes the whole process of development and maintenance easier and faster.

References

[1] M. Kozlenko, I. Lazarovych, V. Tkachuk and V. Vialkova, "Software Demodulation of Weak Radio Signals using Convolutional Neural Network," 2020 IEEE 7th International Conference on Energy Smart Systems (ESS), Kyiv, Ukraine, 2020, pp. 339-342, doi: 10.1109/ESS50319.2020.9160035.

[2] M. Kozlenko, I. Lazarovych, and M. Kuz, "Deep learning approach to signal processing in infocommunications," Proc. 4th International Scientific and Practical Conference on Applied Systems and Technologies in the Information Society, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, pp. 81-82. DOI: 10.5281/zenodo.4482757

[3] Dileep, Kumar & Vuda, Sreenivasa & Yilma, Getinet & Ahmed, Mohammed. (2016). Social Sentimental Analytics using Big Data Tools. 10.1515/9783110469608-027.

[4] Kilinç, Deniz. (2019). A spark-based big data analysis framework for real-time sentiment prediction on streaming data. Software: Practice and Experience. 49. 10.1002/spe.2724.

[5] J. Zaldumbide, R. O. Sinnott, "Identification and Validation of Real-Time Health Events through Social Media," 2015 IEEE International Conference on Data Science and Data Intensive Systems, Pages 9 – 16, doi 10.1109/DSDIS.2015.27.

[6] D. T. Nguyen and J. E. Jung. Real-time event detection for online behavioral analysis of big social data. Future Generation Computer Systems, 2016.

[7] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An architecture for real time analysis of social media text. Proceedings of the workshop on real-time analysis and mining of social streams, 2012

[8] Application of sentimental text analysis to assess public opinion - scientific and technical journal of information technologies, mechanics and optics january–february 2015 vol. 15 no 1 issn 2226-1494

[9] Text analytics accuracy measures. (2021, March 1). Pega. https://community.pega.com/knowledgebase/articles/decision-management/84/text-analytics-accuracy-measures

[10] Sentiment140 dataset with 1.6 million tweets. (2017, September 13). Kaggle. https://www.kaggle.com/kazanova/sentiment140

[11] Twitter Dataset - TrackMyHashtags. (2019). https://www.trackmyhashtag.com/twitter-dataset

[12] Facebook comments Sentiment analysis. (2017). Kaggle. https://www.kaggle.com/mortena/facebook-comments-sentiment-analysis/data

[13] Twitter US Airline Sentiment. (2019, October 16). Kaggle. https://www.kaggle.com/crowdflower/twitter-airline-sentiment