International Journal of Engineering & Technology, 7 (4.40) (2018) 104-107

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Machine Learning in Cloud: Sentiment Analyzing System

Roman Dyussembayev[1], Maryam Shahpasand[1]*

[1]Asia Pacific University of Technology & Innovation (APU)
**Corresponding author E-mail: maryam.shahpasand@staffemail.apu.edu.my*

Abstract

As the number of computer users increases, numerous content has been generated by them. Machine learning as one of the main direction of natural language processing, allows computer systems to extract various information from the generated content. Processing results determine the sentiments of the text to extract the author's emotional evaluation that is expressed in the text. The aim of the project was to develop the Sentiment Analyzing system by using Machine Learning algorithms on cloud-based system. The paper describes the development process of Sentiment Analyzing System in English language. Two Machine Learning algorithms, SVM and Naïve Bayes classifier, have been inspected and Cloud computing used to develop and publish web application. The testing results demonstrate the accuracy of the work in proposed method.

Keywords: *Sentiment analyzing; Machine Learning; Cloud Computing; Natural language processing; Data Science*

1. Introduction

In our time, there is an intensive growth of the Internet adoption. As the number of users increases, the amount of content generated by them increases as well. One of the directions of machine learning, natural language processing, allows computer systems to extract various information from this content. An important task of processing natural language is to determine the sentiments of the text. The task of determining the sentiments of the text is to extract the author's emotional evaluation that is expressed in the text. The data, which went through the system processing, represent author's sentiments and reaction to certain object, person or idea. People, who write blogs, comments, news, have sentiments in their content. Sentiment analysis applications have been widely used in different fields. Researchers have firmly demonstrated the utility of sentiment analyses by successfully correlating changes in opinion expressed in social media with social, political and economic fields [1]. The paper describes a comparison of two machine learning algorithms with the identification of the effect of text preprocessing at the data preparation stage using the messages from Twitter.

2. Literature Review

According to [2], there are three types of the sentimental analysis of text messages:

1. The Analysis of the text according to pre-compiled sentiment dictionaries with the use of linguist analysis. Sentiment dictionaries consist of such elements as words, phrases, patterns, each of which has its own sentiments. The sentiments of the text are determined by the totality of the found emotive vocabulary and is evaluated depending on the amount of positive and negative.

2. Sentiment Analysis of the text using machine learning methods. The text is presented in vector form. Based on the available training sample, the classifier is trained. After this, the text's sentiment is classified.

3. Combination of the first and second approaches. The first approach is rather laborious because of the need to compose sentiment dictionaries, obtain a list of sentiment patterns and develop linguistic analyzers, but it is more flexible. The advantage of this approach is that it allows to see the emotional vocabulary at the sentence level.

According to [3], sentiment Analysis of the text involves the use of one or more methods, each of which has advantages and disadvantages. Methods based on rules and dictionaries, within the framework of these approaches, the text is analyzed based on pre-compiled sentiment dictionaries. But, the process is very long. The main problem is the fact that the same word in different contexts can have different sentiments. Means that for the operation of the system it is required to compose many rules and, most often the systems for analyzing the sentiments of the text are created with reference to a specific subject area. Methods based on graph models. Within the framework of these methods, the text is depicted as a graph based on the assumption that some words have more weight and, therefore, more strongly affect the tonality of the entire text. After ranking the vertices of the graph, words are classified according to the tonality dictionary, where each word is assigned a certain characteristic ("positive", "negative" or "neutral"). The result is calculated as the ratio of the number of words with a positive rating to the number of words with a negative rating.

Machine learning is an idea of the existence of common algorithms that can analyze a set of data without having to write a code specific for the problem. Instead of writing the code, developer pass the data to a common algorithm, and it builds its own logic based on them. For example, one of the types of algorithm is the classification algorithm. It can put data into different groups. The same classification algorithm used to recognize handwritten digits, which can also be used to classify letters for spam and non-spam without changing the line of code. This is the same algorithm, but

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

trained on other data, so the output is a different classification logic.

There are 2 methods based on machine learning according to [3]: supervised and unsupervised. Large data can provide significant assistance in training neural networks, which are also used in the analysis of the sentiments of the text. The principle of the program: it builds a tree with an estimate of the sentiments of each word, each phrase and the whole text. The program understands that changing the order of words changes the key of the text. It can be assumed that this fact provides such a high accuracy in the evaluation of the text and allows to consider neural networks as a promising tool for such analysis.

The work [4] considers the classification of sentiments using machine learning and shows that this approach surpasses simple techniques based on the compilation of dictionaries of frequently used positive and negative words. In the following work [5], the authors describe an algorithm that allows classifying a sentiment using only subjective sentences. Objective sentences, as a rule, do not have sentiments, but create noise in the data. In [6], the authors consider the problem of the fact that a very large number of terms are extracted from the training data when classifying a key. The authors describe methods for selecting the most informative terms and evaluating their sentiments. To eliminate the shortcomings of the approaches discussed above, they are combined. So, in work [7] the method is based on the extracted lexical rules, while training with human participation and machine learning are combined into one algorithm for classifying the key. In another paper [8], researchers from Microsoft suggest ways to reduce the time needed to compose sentiment dictionaries. The result is achieved through the sharing of automatic extraction of informative templates and machine learning.

3. Development Methodology

Azure is the only major cloud platform that Gartner estimates is the industry leader in providing both IaaS and PaaS solutions. This full-featured combination of managed and unmanaged services allows to create, deploy, and manage applications in any way to achieve incredible performance.

To perform the tasks of analytics with predictive analytic using Azure Machine Learning, just need to perform the following steps: upload or import online any current or accumulated data, build and validate the model, and create a web service that uses models to perform quick predictions in real time.

Azure ML is represented by two conceptual components: Experiments and Web Services and one development tool called ML Studio. People who have a Microsoft (Live ID) account can work together in work environment with ML Studio, and they do not even need to pay Azure subscription to work with.

Experiments can be presented as streaming configurations (dataflow) of what developer would like to do with information and models. ML Studio is a web application with the clean interface and works well with all web browsers IE, Firefox and Chrome.

ML Studio starts work by deciding which data sources need to use: the datasets or live data available through the Reader mechanism from a web page, OData, SQL Azure, Microsoft Azure, Hive or Azure blobs. Then, may need to perform some Data Transformation (grouping, renaming columns, merging, eliminating duplicates, discretization operation).

4. Machine learning algorithms

4.1. SVM (Support Vector Machine)

Support vector machine (SVM) have been shown to be highly effective in traditional text categorization, generally outperforming Naïve Bayes [9]. They are large-margin, rather than probabilistic, classifiers, in contrast to Naïve Bayes. In the two-category case, the basic idea behind the training procedure is to find a hyperplane, represented by vector ~w, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. This search corresponds to a constrained optimization problem; letting $c_j \in \{1, -1\}$ (corresponding to positive and negative) be the correct class of document d_j, the solution can be written as:

$$\vec{w} := \sum_j \alpha_j c_j \vec{d}_j, \quad \alpha_j \geq 0 \tag{1}$$

Definition 1: *where the α_j 's is obtained by solving a dual optimization problem. Those ~d_j such that α_j is greater than zero are called support vectors, since they are the only document vectors contributing to ~w. Classification of test instances consists simply of determining which side of ~w's hyperplane they fall on.*

It is convenient to illustrate the idea of the method by the following simple example: points are given on the plane divided into two classes (Fig. 1). Draw a line separating these two classes (the red line in Fig. 1). It belongs to a classification-type algorithm. The algorithm will split the data points using a line. This line should be as far from the nearest data points in each of the two categories. Further, all new points (not from the training sample) are automatically classified as follows:

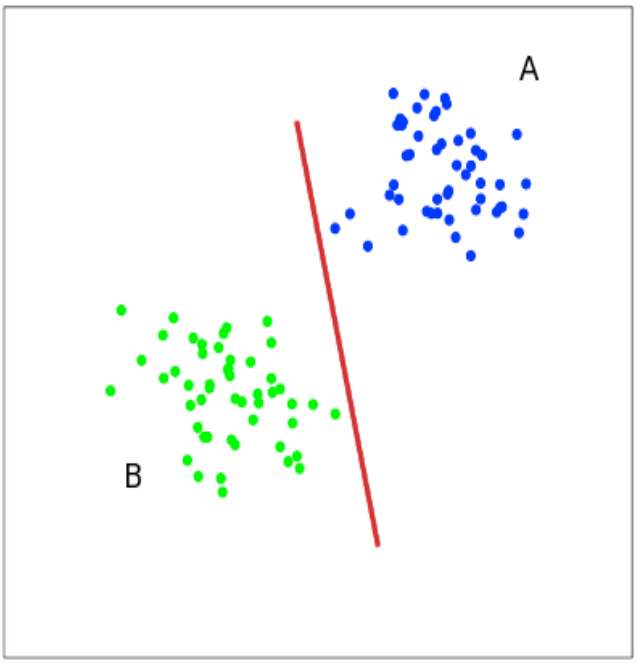

Fig. 1: The point above the line is in class A, The point below the line is in class B.

4.2. Naïve Bayes classifier

The Naïve Bayesian classifier is one example of the use of vector analysis methods. Naïve Bayes is one of the most frequently used classifiers, because of comparative simplicity in implementation

and testing. At the same time, the Naïve Bayesian classifier demonstrates not the worst results, in comparison with other, more complex classifiers. The Naïve Bayesian classifier is based on the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

Definition 2: *P(A|B) is the posterior probability of class (target) given predictor (attribute). P(A) is the prior probability of class. P(B|A) is the likelihood which is the probability of predictor given class. P(B) is the prior probability of predictor.*

A Naïve Bayesian classifier is a family of classification algorithms that accept one assumption: Each parameter of the classified data is considered independently of other class parameters. Parameters are called independent when the value of one parameter does not affect the second one. The theorem allows us to predict a class based on a set of parameters using probability.

$$P(\text{Class A}|\text{Feature 1}, \text{Feature 2}) = \quad (3)$$

$$\frac{P(\text{Feature 1}|\text{Class A}) * P(\text{Feature 2}|\text{Class A}) * P(\text{Class A})}{P(\text{Feature 1}) * P(\text{Feature 2})}$$

Definition 3: *The equation finds the probability of class A, based on parameters 1 and 2. The probability of class A on the basis of parameters 1 and 2 is a fraction.*

5. Problem context

The first problem of sentiment analysis is the Challenges in Multilingual Sentiment Analysis [10, 11]. This is due to the complexity of compilation and the large amount of data obtained. Work on multilingual sentiment analysis has mainly addressed mapping sentiment resources from English into morphologically complex languages.

The second problem that defines in [10] is Sentiment in Figurative Expressions. The data in the dictionaries are divided into positive and negative context. System can analyze many sources, but some phrases and words would sound innocuous. In fact, their real meaning will be unnatural and illegal. Figurative expressions in text, by definition, are not compositional. That is, their meaning cannot fully be derived from the meaning of their components in isolation. There is a growing interest in detecting figurative language, especially irony and sarcasm. Another problem is the regular updating of the dictionary [12, 13].

6. Implementation

6.1. Training dataset

The data used in this experiment is Sentiment140 dataset, a publicly available data set created by three graduate students at Stanford University [14, 15]. The data comprises approximately 1,600,000 automatically annotated tweets. 160,000 rows were extracted to use as the dataset. Positive and negative tweets are equally distributed.

6.2. Cloud

The Azure Machine Learning is a cloud-based service for performing predictive analytics. The service is represented by two components: Azure ML Studio and Azure ML web services. Microsoft Azure Machine Learning Studio is a collaborative, drag-and-drop tool that developers can use to build, test, and deploy predictive analytics solutions on data. Machine Learning Studio publishes models as web services that can easily be consumed by custom apps or BI tools such as Excel. Projects in Azure ML Studio are called experiments. Experiments can be represented as streaming configurations (data-flow) of what the developer would like to do with information and models. As an Azure ML data researcher, developer focus on experiments and can spend all the time in ML Studio, only doing re-engineering experiments, changing parameters, algorithms, validation criteria, periodically making changes to data, and so on. The interface looks clean, nice and works well not only in IE, but also in Firefox and Chrome. In addition, the developer can use R inside Azure ML. Azure ML contains enough packages for comfortable work with R. According to the research above, Azure ML is suitable environment for this project since the project is cloud based and supports R. Cloud Infrastructure can be used on any OS or browser.

6.3. Testing

The Accuracy test is calculated as the ratio of all successful predictions to the total number of elements in the set: (True Positive + False Negative) / Total numbers. For 80,000 positive and 80,000 negative tweets in dataset (70% - training dataset, 30% - testing dataset). Each training and test sets contains the same number of positive and negative tweets. Precision is evaluated by checking how consistent results are when measurements are repeated. Precise values differ from each other because of random error, which is a form of observational error.

Table 1: Accuracy and precision results

Text preprocessing	**Yes**		**No**	
Algorithm	SVM	NB	SVM	NB
Accuracy	76,7%	77%	80,2%	88,6%
Precision	75%	76%	79,4%	88,1%

Precision refers to the closeness of two or more measurements to each other. Precision and recall are metrics that are used for evaluating information extraction algorithms. The accuracy of a system within a class is the proportion of documents belonging to a given class relative to all documents that the system has assigned to this class. The completeness of the system is the proportion of documents found by the classifier belonging to a class relative to all documents of this class in the test set. Precision is independent of accuracy. System can be precise but inaccurate, also be accurate but imprecise.

7. Conclusion

Problem regarding to [3] can be solved or at least reduced by implementing system that has been trained on the same data source as the input data. The research shows a significant increase in accuracy when specific tool is used for specific data or task. The creation of a universal system that can accurately determine the data from any source is so far unlikely. For the second problem, additional type of data in training dictionary can lead to the overfitting. Overfitting is the production of an analysis that corresponds too closely or exactly to a set of data and may therefore fail to fit additional data or predict future observations reliably. Specific training dataset for the specific data analyzation produce enough features to ensure that such expressions do not significantly affect the result of the analysis.

As previously thought Messages in social networks contain information that is not needed for the sentiments analysis: links, user names and hashtags. This research has shown that for the analysis of tweets it is necessary to consider all kinds of data without preliminary preprocessing of the text. It is also not recommended to preprocessing the training dictionary when the system is being used only for specific data recourse. The researcher suggests implementing unsupervised model for text message normalization

explained in [17]; this method could provide more accurate pre-processing.

To improve the estimation of the efficiency of algorithms, it is supposed that it should be supplemented with elements of linguistic analysis. The research [12, 16] shows that the best results are achieved by combining linguistic and statistical approaches. More advanced methods of deep learning predicting behavior and game theory discussed in [18, 19]. To classify sentiments for multi-language, the research [20] provides a solution for multi-language sentiment classification.

References

[1] Bollen J., Pepe A., Mao H., (2011). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. Barcelona, *Fifth International AAAI Conference* on Weblogs.

[2] Yussupova, Bogdanova, Boyko, 2014. Applying of Sentiment Analysis for Texts in Russian. Ufa, The Second International Conference on Advances in Information Mining and Management.

[3] Marouane B., 2017. Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. Procedia Computer Science, Volume 113, pp. 65-72.

[4] Pang B., Lee L., 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization. Ithaca, NY 14853-7501: Department of Computer Science, Cor-nell University.

[5] Pang B., Lee L., 2008. Opinion Mining and Sentiment. 2 ed. NY USA: Computer Science Department, Cornell University.

[6] Mehdi A., Seyedamin P., Saied S., Elizabeth D., Krys K. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. CoRR, Volume: abs/1707.02919.

[7] Prabowo R., Thelwall M., 2009. Sentiment analysis: A combined approach. School of Computing and Information Technology, 143-157(2), pp. 1-21.

[8] König A., Brill E., 2016. Reducing the Human Overhead in Text Categorization. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ISBN: 978-1-60558-193-4, pp. 274-282.

[9] Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. *ECML 1998: Machine Learning: ECML-98*, Issue 1398, pp. 137-142.

[10] Mohammad, S. M., 2015. *Challenges in Sentiment Analysis.* Montreal, Canada: National Research Council Canada.

[11] Rentoumi, Petrakis, Klenner, Vouros, Karkaletsis, 2010. United we stand: improving sentiment analysis by joining machine learning and rule-based methods. Malta, 7th International Conference on Language Resources and Evaluation.

[12] Ali H., Sana M., Ahmad K., Shahaboddin S. 2018, "Machine Learning-Based Sentiment Analysis for Twitter Accounts", Mathematical and Computational Applications 2018, 23(1), 11

[13] Devendra S., Manzil Z., Ruslan S. 2018, "Investigating the Working of Text Classifiers", CoRR, abs/1801.06261

[14] Alec Go, Richa Bhayani, Lei Huang, 2009. Twitter sentiment classification using distant supervision. pp. 1-6.

[15] Monireh E., Amir H., Amit S. 2017, "On the Challenges of Sentiment Analysis for Dynamic Events
", IEEE Intelligent Systems, Volume: 32, Issue: 5, pp:70-75

[16] V.Uma Ramya, K. Thirupathi Rao 2018, "Sentiment Analysis of Movie Review using Machine Learning Techniques", International Journal of Engineering & Technology, 7 (2.7) 676-681

[17] P. Cook, S. Stevenson 2009, "An unsupervised model for text message normalization", NAACL HLT 2009 Workshop on Computational Approaches to Linguistic Creativity, 71-78, Boulder, Colorado.

[18] J. Wright, K. Leyton-Brown 2017, "Predicting Human Behavior in Unrepeated, Simultaneous-Move Games", Games Theory and Economic Behavior (GEB), volume 106, pp. 16-37

[19] J. Hartford, J. Wright, K. Leyton-Brown 2016, "Deep Learning for Predicting Human Strategic Behavior", Oral presentation at Conference on Neural Information Processing Systems (NIPS)

[20] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, M. Jaggi 2017, "Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification" In Proceedings of the 26th International Conference on World Wide Web