Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 329–335

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner

Vishal Vyas*[a], V.Uma[b]

[a] *Vishal Vyas, Department Of Computer Science, Pondicherry University, Puducherry- 605014, India*
[b] *V.Uma, Department Of Computer Science, Pondicherry University, Puducherry- 605014, India*

Abstract

The online content includes conversation in social networking websites, tweets, blogs and various forums discussing occasions, people and everything which exist in this world. With the huge growth of the online content, high rated information is achievable. Customer feedbacks help organizations to improve their services by rectifying their drawbacks. Manual analysis of information in the era of big data would be a cumbersome task. Countless tools are available for mining information/sentiment from World Wide Web but the choice of tool is the biggest problem at present time. One should have an idea about the framework to evaluate tools. With the evaluation structure for the tools, this paper will compare twenty tools for text analysis with respect to their applications and extension availability for Sentiment Analysis (SA). An experiment is performed in Rapid Miner to derive Sentiment from tweets and accuracies of different algorithms are compared to find out the best performing one.

© 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications

Keywords: Text mining; Sentiment analysis; Twitter; Rapid Miner

1. Introduction

The errand of Sentiment examination in social media content is troublesome because of variability and intricacy of dialect articulation. The not only complexity that makes the task difficult but also the availability of real-time content in a huge quantity is an issue. Manual sentiment analysis is an unmanageable task hence an efficient tool or intelligent system is the need of the hour. Organisations need an intelligent tool to formulate an excellent insight from web-based social networking information. Yet the analysis provided by the vast majority of the tools does not meet their expectations. Tools which monitor online content in social networking websites empower ventures to analyse clients

* Corresponding author. Tel.: +91-7598593973.
E-mail address: vyasvishaluni@gmail.com

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.044

supposition, inquiries at real time in a profoundly versatile manner and many more. Availability of tools to monitor online content has quickly expanded in recent years. Enterprises are confronted with the troublesome assignment of picking the right tool according to their necessity. There are numerous vendors who add new components to empower existing tool which helps clients to transform online networking information into significant knowledge. Machine learning (ML) algorithms such as support vector machine[1], decision tree[2], naive bayes[3] and many more acts as an important component while designing a tool. Various supervised[4] and unsupervised learning[5] machine learning classifications are used according to the tool's purpose. This results in a variety of tools but the user gets confused during the selection of a tool to empower their organisation with social media intelligence. These tools offer means for observing the online networking users, dissecting and measuring their sentiments toward a brand, prepare multifarious insight which ultimately helps organizations to improve or enhance their services. Concepts of choosing a tool for text analysis/SA are discussed in this paper. These concepts to evaluate tools comprises of seven areas which help in the selection of the best tool. Considering these concepts, any organisation approaching a new tool not only gets a right tool for getting insight from online media but also benefits the business to reach new heights [6]. Concepts such as Current offering, Workflow management, Market presence, Analysis, Engagement, Strategy and Influence are collectively called as a framework for evaluating a tool. Which tool is beneficial for an organization? To answer this question one must have the knowledge of the framework to evaluate intelligent tools. Section 2 of this paper provides a detailed explanation of a framework to evaluate a tool. In this section, all the necessary concepts which should be taken care of before approaching a new tool for sentiment analysis are clearly explained. In section 3 various tools are categorized with respect to parameters such as application, web sources and extensions available for Sentiment Analysis. The Categorization table provided will help researchers in getting a brief insight of available tools for sentiment analysis. Section 4 will present an experiment with twitter dataset using Rapid Miner to deduce sentiment (positive/negative).

Nomenclature

SA - Sentiment analysis
SVM - Support Vector Machine
OSS - Open source software
NLP - Natural language processing

2. Framework to evaluate a tool

Choosing a tool from the market which at present has nearly two hundred tools is a difficult task. Which tool does what is identifiable but the question is that what thing/concepts should be taken care of while approaching a new tool. How it is different from other available tools which also perform the same task can only be known if certain concepts are considered. Concepts on which one should focus while approaching a new tool are Analysis, Current offering, Engagement, Influence, Strategy, Market presence and Workflow management. Considering these concepts one can identify an efficient tool which can help the enterprise. These seven concepts form a framework to evaluate a tool. In this list Current Offering, Strategy and Market presence are suggested by [Hofer-Shall et al][7]. Considering this list as insufficient [Stavrakantonakis et al][6] added more areas and made a detailed framework for the evaluation of tools.

2.1. Current Offering

Satisfaction of a customer should be the top priority of an enterprise. By satisfaction, it means the tools ability to record social media information, its examination and incorporation with the customers requirement. A customer should assess how well an enterprise set up and actualizes enquiries propose new extensions and associated social information with business statistics.

2.2. Workflow Management

It alludes to the way toward reacting to online media streams. Tools ability to avoid twofold responses and duplicate suggestions is vital for the productivity of an organisation. Tools compatibility with other tools plays an important role.

2.3. Market Presence

The dynamic client base of an enterprise gives a broad picture of its popularity in the market. The Financial transactions of an enterprise tell its presence. How much area a particular enterprise is covering shows the reliability of the services it is offering. A maintained market presence by any enterprise is not easily attained and it shows its popularity.

2.4. Analysis

The chosen intelligent tool ought to have the capacity to assemble information from many sources and in various structures such as posts, reviews, Audio and Video. Tools ability to create a database to store the captured information and its visualization techniques makes it different from others. Capability to remove the undesirable information such as duplicate words and process the necessary information to produce a meaningful insight of data is the most look out parameter in a tool.

2.5. Engagement

Many tools nowadays have the feature to work in real time environment. Its ability to participate in real time conversation by analysing the customer feedbacks improves the business and enhances the customers faith in the organisation. The decrease in the response time to handle the queries of customers ultimately help organisations to welcome more clients.

2.6. Strategy

Strategy is the master plan to achieve an overall aim. It represents how well an enterprise is focused on addressing the customer feedbacks and supporting customer by addressing customer enquiries. The certainty of an enterprise to incorporate new technologies and efforts by conducting surveys to consider customer feedbacks makes them different from other competitors.

2.7. Influence

Influence alludes to those posts (consider social media) that affect individuals. A tool for text analysis must define the initiation of the post and count of people following the same. Tools ability to extract features from the text does this job and identifying the intensity of sentiment differentiates people involve in such activities. This concept protects an organisation from data theft, information leakage and many more harmful activities.

All the tools listed in Table 1 satisfies above criteria. These Twenty tools are considered to be the best tools available in the market. Some of these are not free software but a tool like Rapid Miner is an open-source software (OSS). Being an OSS it benefits researchers to perform the experiment on it. There are many OSS in the market but the user interface and workflow of Rapid Miner [8] makes it different from other tools. To infer the sentiment from tweets and to compare the accuracy of different algorithms Rapid Miner is used.

3. Categorization of tools for Sentiment Analysis

A large number of tools are available for text analysis/ Opinion mining/ Sentiment Analysis. Many tools which are designed for text analysis come with a feature of adding extensions to enhance primary features of the existing

tool. Tools along with details about their application and extensions are provided in table 1. The table consists of four columns. The first column represents a list of twenty tools for text analysis/SA. The second column represents the various applications of the tools. The third column represents the web sources and the fourth column represents the available extension for Sentiment Analysis.

Table 1. Categorization of Text Analysis tools

Tool Name	Application	Web Source	Extention for Sentiment Analysis
Lexalytics	Natural Language Processing (NLP) , Text Analysis	https://www.lexalytics.com/ technology/sentiment	Saliance
IBM Watson Alchemy API	NLP, Text Analytics, Content Recommendation	https://www.ibm.com/watson/ developercloud/alchemy-language.html	AlchemyAPI
Provalis Research Analytics Software	Text analysis, Content Analysis	https://provalisresearch.com/	WordSat
SAS Text Miner	Text analysis, Ontology, Sentiment Analysis, NLP	https://www.sas.com	Nil
Sysomos	Social media monitoring, Text Analysis	https://sysomos.com	Media Analysis Platform (MAP)
Expert system	Semantic Search, NLP, Conent Analysis	www.expertsystem.com	Cogito
Rapid miner	Text mining, Social media analysis, Market Search	https://rapidminer.com	Setiment Analysis
Clarabridge	NLP, Text Analytics, Social media Analysis, Sentiment Analysis	www.clarabridge.com/ text-analytics	Nil
Luminoso	Text analysis	https://luminoso.com	Luminoso Compass
Bitext	Sentiment Analysis, Concept Extraction, Text Analysis	https://www.bitext.com/	Nil
Etuma	Social media monitoring, Sentiment Analysis	https://www.etuma.com/	Nil
Synapsify	Social media monitoring, Text mining	www.gosynapsify.com	Snapify core API
Medallia	Social media monitoring, Text mining	www.medallia.com	Nil
Abzooba	Social media monitoring, Text Analysis	www.abzooba.com/	XPRESSOInsight
General Sentiment	Sentiment Analysis, Text mining, Social media analytics	www.generalsentiment.com	Nil
Semantria	Text analysis by API and Excel plugin, Sentiment Analysis	https://semantria.readme.io/	Nil
VisualText	NLP, Text Analytics	www.textanalysis.com	Nil
Buzzlogix	Text analysis, Sentiment Analysis, Social Media Monitoring	https://buzzlogix.com	Nil
Averbis	Text analytics, Information Discovery	https://averbis.com/en/	Nil
AYLIEN	Text analysis, NLP, Concept Extraction	aylien.com	Nil

4. Analysing tweets using Rapid Miner

Problem Statement: To determine the polarity of tweets i.e positive or negative.
Classification Models used: Support Vector Machine (SVM), Decision tree and Naive Bayes.
Training data: 400+ tweets.
The two attributes in training data are tweets and sentiment. Nearly 450 tweets were captured from Twitter and were manually labelled with the sentiment positive/negative to train the model. Trained model is then used for identifying sentiment in the test set. Rapid Miner contains many operators, each for a specific requirement. In this experiment read excel operator is used to input raw data. The raw data is captured from twitter and stored in a spreadsheet having 450 rows and the tweets and sentiment as two attributes. Data received from twitter contains undesired words, hence preprocessing of data is very important [9]. For preprocessing, an operator named process document from data is used and various steps (by adding operators) are performed. Five operators are used to complete the said stage. The first operator transforms the document into lower cases. Secondly, by using Tokenization operator the text in document splits into a sequence of tokens. There are many unwanted tokens such as is, am, are etc which don't have any specific sentiment, hence these are removed by using filter stop word (English), which is the third operator. Fourthly, Filter tokens by length are used to remove word according to their length. The length of the word is customized. The author has filtered those word which has minimum length 3 and maximum length of 20. The fifth stage of preprocessing contains an operator named stem that transforms words in the form of tokens to their base form. For stemming Porter algorithm [10] is used. Split validation technique is used to train SVM, Decision tree and Naive Bayes separately. Entire training dataset is split (split ratio = .66) into training and test dataset. With the split ratio of .66, it means 2/3rd of data is used for training and remaining is used for testing. The test dataset is stored in a spread sheet. After preprocessing this dataset is given as an input to the trained model where sentiments in this document are identified. Calculation of parameters such as Precision, Recall and Accuracy is done to choose the best model to use in this experiment. Fig 1 shows the performance of three models.

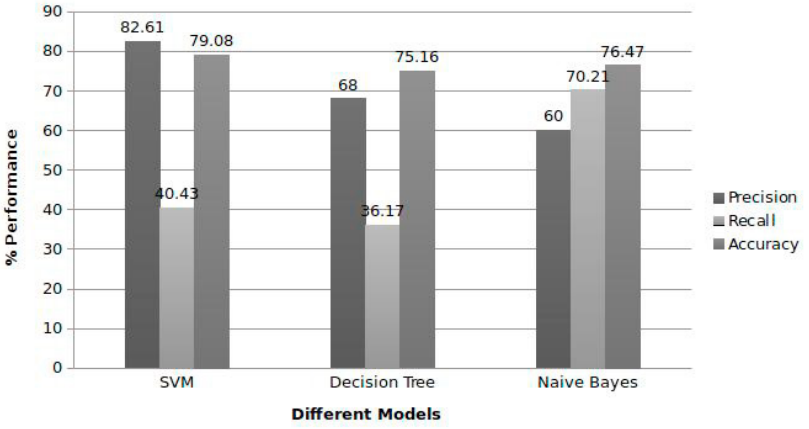

Fig. 1. Performance of different models with 450 training datasets

It is found that SVM shows high precision and accuracy as compared to other models with similar training data set. What will be the accuracy of SVM, Decision tree and Naive Bayes in different size of training samples? The answer to this question is in Fig 2. It shows some glitches in other models whereas SVM shows a gradual increase in the accuracy. To inspect the accuracy behaviour of SVM, sample size of 200 tweets and 420 tweets are introduced and it is found that accuracy of SVM may vary with the increase in the size of training data. Fig 3 shows the variation in accuracy of SVM with respect to the training samples. But 79.08% accuracy of SVM is still the highest among other two models. Hence in validation and testing of tweets dataset, SVM is used in this experiment.

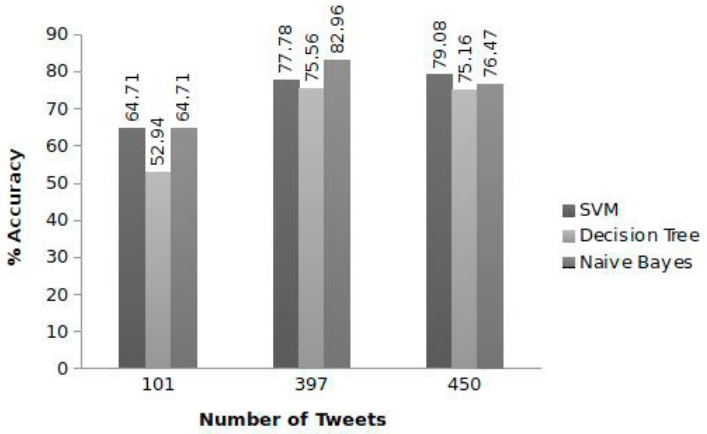

Fig. 2. Accuracy of three models with respect to different sample size of training dataset

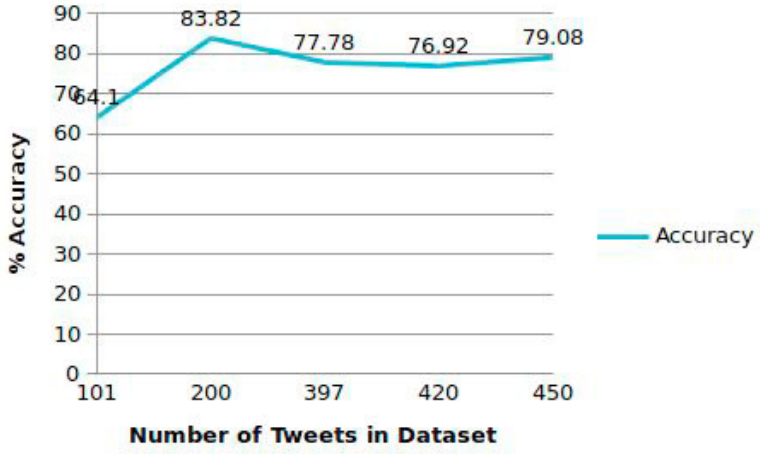

Fig. 3. Accuracy of SVM with different sample size of training dataset

5. Result

With the training data set of 450 tweets, SVM has shown 82.61% precision, 40.23% Recall and 79.08% Accuracy. The performance of SVM is found to be better than other models. Fig 4 shows the function value is assigned to positive and negative tweets during the validation process. Here, it is identified that function value depends on the word occurrence in a tweet and Term frequency id (TF-idF)[11] associated with the word.

6. Conclusion

Information is achievable from over flooded content in World Wide Web. A variety of tools are available these days to get an insight from online content. After the evaluation of tool on all areas discussed in this paper, it is possible to get a right intelligent tool for any organization and for a specific purpose. In this paper, the efficiency of the Rapid miner is identified by using it to deduce sentiments (positive/negative) from tweets. Its easiness and portability make it different from other available tools. We explored many features of Rapid miner while analysing sentiment from

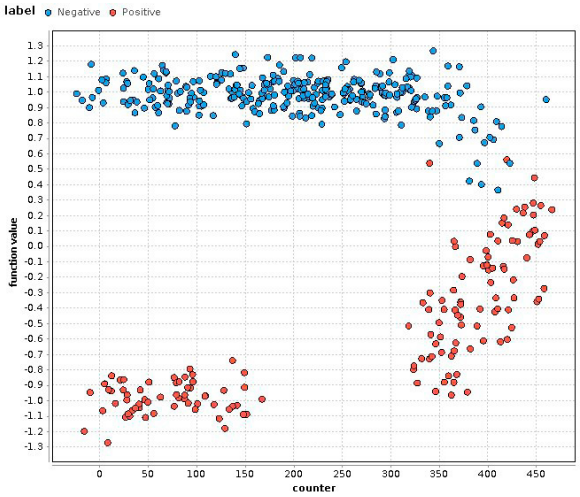

Fig. 4. Behaviour of training dataset

tweets. We used SVM, Decision tree and Naive Bayes during validation of training data and found SVM performs expertly. With the training dataset of 450 tweets, SVM achieves 79.08% accuracy in identifying sentiments.

References

[1] Tong, S., Koller, D. (2001)"Support vector machine active learning with applications to text classification."*Journal of machine learning research* **2**(Nov) 45–66

[2] Rokach, L., Maimon, O.(2014)"Data mining with decision trees: theory and applications." *World scientific*

[3] Zhang, H. (2004)"The optimality of naive bayes." *AA* **1**(2) 3

[4] Hastie, T., Tibshirani, R., Friedman, J. (2009)"Overview of supervised learning." *In: The elements of statistical learning.* Springer 9–41

[5] Aggarwal, C.C., Zhai, C.(2012)" Mining text data." *Springer Science & Business Media*

[6] Stavrakantonakis, I., Gagiu, A.E., Kasper, H., Toma, I., Thalhammer, A."An approach for evaluation of social media monitoring tools." *Common Value Management* **52**(1) (2012) 52–64

[7] Hofer-Shall, Z.(2010)"The forrester wave¢: Listening platforms, q3". *Forrester Research*

[8] Santhanakumar, M., COLUMBUS, C.C.(2015)" Web usage based analysis of web pages using rapidminer." *Wseas Transactions On Computers* **14**

[9] Feldman, R., Sanger, J.(2007)"The text mining handbook: advanced approaches in analyzing unstructured data." *Cambridge university press*

[10] Willett, P.(2006)"The porter stemming algorithm: then and now." *Program* **40**(3) 219–223

[11] Ramos, J., et al." Using tf-idf to determine word relevance in document queries." *In: Proceedings of the first instructional conference on machine learning.* Volume 242. (2003) 133–142