International Journal of Industrial Engineering & Production Research March 2020 Vol. 31, No. 2: 317-322
DOI: 10.22068/ijiepr.31.2.317

RESEARCH PAPER

An Empirical Review of Challenges of and Approaches to Sentiment Analysis

P. Subbaraju*[1], K. Chandra Sekhar[2], P. R. S. S. V. Raju[3], K. Satyanarayana Raju[4] & M. K. S. Varma[5]

Received 18 April 2020; Revised 28 April 2020; Accepted 3 May 2020; Published online 30 June 2020

ABSTRACT
Nowadays, data are food for the digital world. The main rich sources of data are generated by social networks such as Twitter, Facebook, Instagram, and LinkedIn. The data generated through micro-blogging services play a vital role in business intelligence such as product reviews, movie reviews, and election results prediction by social media data analysis. Sentiment Analysis (SA) is the key method of predicting netizens' emotions behind the text expressed in social media. The main objective of this survey is to shed light on the tools and techniques used in Sentiment Analysis and the relevant fields in brief detail.

KEYWORDS: *Business intelligence; Data generation; Sentiment analysis; Social media data analysis.*

1. Introduction

Social media have drastically changed the lifestyle of the public by sharing and expressing their ideas and opinions in the form of texts, images, and videos through websites like Twitter, Facebook, and Instagram. Social media represent one of the best sources where one can find business intelligence on product-based companies and reach a partial assessment based on customer reviews or feedback expressed; platforms like these are favorable enough to analyze public sentiment. Sentiment Analysis (SA) is an approach to determining the emotional tone behind the body of a text. SA is also known as text mining or opinion mining. SA process has gained popularity in recent years and it involves categorizing text data, which is generated from social networking websites. Classifying the public opinions is a difficult task because every person may voice his/her unique positive, negative, or neutral views about a certain product. In addition, shared opinions are often seen in the unstructured format. In order to analyze unstructured text data and retrieve information on people's sentiment, machine learning algorithms should be used with new techniques. In sentiment analysis, major new challenges are 1) sarcasm detection, 2) troll detection, 3) abuse detection/Cyber-bullying, 4) spam detection, and 5) behavior analysis based on text data in social media.

This paper is organized as follows: Sections 2 and 3 present SA workflow explanation and a survey of related literature, respectively.

Sections 4, 5, 6 discuss and offer approaches to sentiment analysis, tools and metrics used in SA, and conclusion, respectively.

2. Workflow of SA Process

A. *Data collection*: The primary step involves the collection of data, generated throughout blogs, forums, and social media like Twitter and LinkedIn. Information may come in different formats like unstructured, semi-structured, different slangs, languages, and contexts.

B. *Text Preparation*: Before starting the analysis, the text data need to be prepared. The following steps are performed to get the right polarity of the text.

a) Removing numbers,
b) Removing URLs and links,

* **Corresponding author:** ***P. Subbaraju***
raju.pericherla74@pec.edu

1. *Department of IT, SRKREC, Bhimavaram, A.P, India.*
2. *Department of IT, SRKREC, Bhimavaram, A.P, India.*
3. *Department of IT, SRKREC, Bhimavaram, A.P, India.*
4. *Department of IT, SRKREC, Bhimavaram, A.P, India.*
5. *Department of IT, SRKREC, Bhimavaram, A.P, India.*

c) Removing stop words,
d) Stemming words, and
e) Removing punctuations.

C. *Sentiment Detection*: This focuses on subjective information such as views, feelings, etc. rather than objective information.

D. *Sentiment classification*: Classification can be done based on the following methods:

1) ML (Machine learning) based approach,
2) Lexicon based approach.

E. *Presentation of OUTPUT*: Charts, lines, and graphs containing the analyzed information are used to display data visually. Fig 1. displays the steps involved in the SA process.

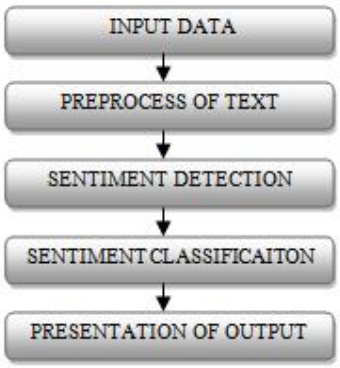

Fig. 1. Process of Sentiment analysis

3. Literature Survey

This section covers discussion on the challenges of SA mentioned in the introduction part.

Sarcasm Detection:

The word Sarcasm is derived from French word sarcasmor which is used to tone down and provide a cover for hate or dislike. In 2019, Le Hoang Son et al. [1] applied sAtt-BLSTM convNet and tested it against unbalanced and balanced datasets. The model excels with classification accuracy of 93.71% for the randomly collected tweets and 97.87% for semEval dataset on sarcasm. In 2018, Rajesh Basask et al. [2] automated public shaming detection in Twitter by classifying it into six types: whataboutery, sarcasm, religious, passing judgment, comparison, abusive. Blockshame web application is designed to protect the twitter users from shamers and plans to investigate multiple comments in a single event for both non-shame and shaming categories with large datasets. In 2016, Mondher Bouazizi et al. [3] proposed a new approach to identify sarcasm based on patterns. The latest pattern-based technique ensures 83.1% accuracy and 91.1% precision. In future works, one can draw on the output of the present research to promote the performance of opinion mining and sentiment analysis.

Troll Detection:

Trolling can be defined as upsetting the public by sharing inflammatory or off-topic messages in online social media. In 2018, Mohd Fazil et al. [4] presented a hybrid method to detect automated spam amalgamating community-based characteristics like metadata, total elements, and essential features in an interactive-based manner. The approach newness lies in the user characterization along with followers' activities and it considers nineteen features with three learning classifiers. In 2016, Jorge De-La-Pena-Sordo [5] applied a new approach to identify trolling comments on social media websites and extract a combination of opinion, syntactic, and statistical features from the user comments. In 2015, Patxi Galan Garcia et al. [6] proposed an advanced technique to completely detect all false profiles on Twitter by analyzing comments generated by the users. SMO-PolyKernel method is performed with 68.47% accuracy and 0.96 AUC.

Abuse/Cyber-bullying Detection:

Cyber-bullying is a type of violence that takes place across digital gadgets in terms of mobiles, smart phones, desktop computers, tablets, etc. It includes SMS, text, social media, and forums. In 2018, Gille Jacobs et al. [7] proposed automatic cyber-bullying detection using SVM by performing binary classification methods. The classifiers return an F1 score of 64% and 61% changing the hyper parameters. In 2018, Debajan Mahata et al. 8] developed a classifier for detecting mentions of personal intake of drugs in tweets. In 2018, Han Hu et al. [9] used target data collection and two-stage annotation techniques to generate annotated datasets. The authors performed detecting drug abuse risk in tweets with accuracy of 86.53% and, also, 88.51% of recollect with 86.63% F1 score. In 2017, Nhathai Phan et al. [10] incorporated essential techniques of Machine Learning (ML) to significantly detect different tweets of drug abuse.

Spam Detection:
Spam refers to sending unwanted or unrequested messages in online forums, blogs, and emails. In 2019, Manisha Sharma et al. [11] introduced an advanced hybrid model composed of neural networks and Long-Short-Term-Memory (LSTM). The novel model outperformed conventional models in terms of accuracy. In 2018, Srikant Madisetty et al. [12] developed a model using neural networks and deep learning techniques. The proposed method enures 0.95% accuracy on HSPAM14 dataset. In 2018, Isa Inuwa-Dutse et al. [13] proposed an approach to optimizing historical tweets, which are available on Twitter for a short while. The authors observed that at least 12 tweets per day belong to an automated spam account. For future works, the effect of an increase in the maximum length of tweets on spamming activities should be considered.

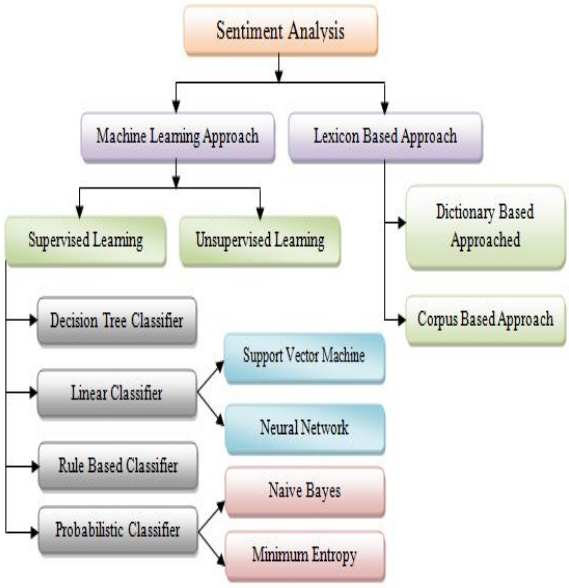

Fig. 2. Techniques of Sentiment Analysis

Personality Prediction in Social media:
In 2018, Bo Xu et al. [14] studied the personality behavior of each Facebook user regarding various features and also meticulously compared four essential machine learning techniques. Typically, the personality prediction builds upon the XGBoost classifier which outperforms with the more than 74.2% prediction accuracy. In 2018, Yulia Tyshchuk et al. [15] developed behavioral patterns in the context of important events and developed a framework for particular behavioral patterns shown in social media and measuring those patterns of behavior. In 2018, Wenli Ji et al. [16] proposed a method for the gender-based identification of reposting behaviors in online social networks. The method outperforms with an accuracy rate of 86.7%.

4. Machine Learning Approaches

A) ***Supervised Learning:***
Supervised Learning is extensively used to train a machine using labeled data, which is the best approach in classification and is widely used for classifying sentiment polarity with precise results. Different techniques of supervised learning to improve the accuracy of classification are given below.

B) ***Naïve Bayes:*** it is a purely classification technique that depends on probability, i.e., Bayes theorem. It is simpler to build and particularly used in large datasets. It is a well-known algorithm for classification problems

B) Decision Tree: It is the most popular algorithm in data mining, statistics, and machine learning algorithm. It is mostly used in classification. One advantage of using the decision tree is minor data cleaning requirement. The major limitation of the decision tree is the overfitting problem.

C) ***Support Vector Machine:*** It is a very high-performance technique, but makes small tuning and is mostly used in binary classification problems. SVMs involves advertising and gender finding in terms of images to solve a large number of image

classifications. Fig. 3 shows SVMs in a two-dimensional view.

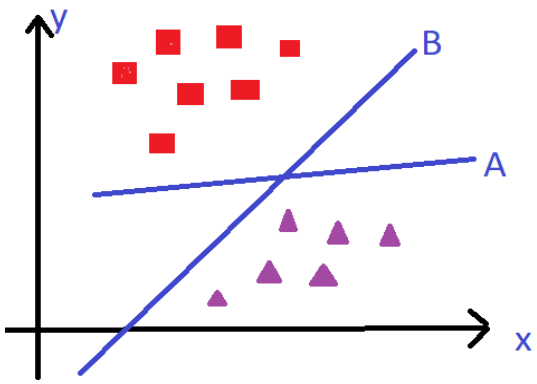

Fig. 3. Support vector machine

D) ***Unsupervised Learning:***
Unsupervised Learning is used to train the model without reference to labels and known datasets. The essential task is to make the non-sorting datasets as group regarding multiple patterns, but without training labeled data. Clustering and association come under unsupervised learning techniques.

E) ***k-means clustering algorithm:***
means clustering algorithm is an iterative approach trying to separate each dataset into k subgroups, each called a cluster; here, k is a variable of total subgroups of cluster. Each cluster refers to a collection of data points with close similarities.

F) ***Hierarchal clustering algorithm***: It is a method of cluster strategy that creates a cluster of hierarchies. There are two types of hierarchal clustering categories: 1) Agglomerative (bottom-up) clustering and 2) Divisive (top-down) clustering.

G) ***Lexicon based Approach:***
In this approach, each word is annotated by a polarity score to detect sentiment of the expression. This method does not require any training dataset. The major drawback is that it cannot include many words and expressions to sentiment lexicons.

5. Sentiment Analysis Tools

Nowadays, several analysis tools are available in the open-source market. They have been used for NLP (natural language processing). This section presents some of the popular tools used for sentiment analysis.

Tab. 1. Sentiment analysis review

Name of the tool	Description
NTLK	NLTK stands for Natural Language Tool Kit. It contains most powerful library packages for preprocessing take tokenization, Stemming, lemmatization, word count, etc.
Open NLP	Open NLP supports the most frequent NLP tasks like sentence separation, POS tagging, parsing, language detection, chunking
WEKA	This tool includes many machine-learning techniques for data mining tasks. It is implemented by JAVA programming language. It supports many algorithms like clustering, classification, and linear regression for analysis.
IBM Watson	It supports advanced text analytics systems in 13 languages. Users can extract keywords, entities, and concepts.
Open Text	It identifies and evaluates subject patterns.

6. Metrics Used in Sentiment Analysis

The general metrics used to evaluate sentiment analysis are accuracy, precision, recall, and F1-score. Depending on the data, an appropriate metric should be considered.

Accuracy: It predicts each observation ratio with high sensitivity.

Accuracy = (TN + TP) / (TP + TN + FP + FN).

Precision (P): It is the ratio of true positives to sum of true and false positives.

Precision = TP / (TP + FP).

Recall (R): It is also called as sensitivity, the ratio of true positives to the sum of true positives and false negatives.

Recall = TP / (TP+ FN).

F1_score: F1_score expresses the balance between recall and precision.
F_score = 2 * (P *R) / (P +R).
Accuracy, precision, recall, and F1-score for the above information are given in Table 1.
Accuracy = (TN + TP) / (TP + TN + FP + FN)= (5 + 13) / (5+7+1+13)
Precision (P) = TP / (TP + FP) = (5) / (5 +7)

Recall(R) = TP / (TP+ FN) = (5) / (5 + 1)

F_score = 2 * (P *R) / (P +R). = 2 * 0.4 * 0.8/ (0.4 + 0.8)

Tab. 2. Confusion matrix

	Actual Positives	Actual Negatives
Positive predictions	5 (TP)	7(FP)
Negative predictions	1 (FN)	13(TN)

7. Conclusion

This study aimed to present an overview of recent changes in and updates of sentiment analysis and classification methods. Many of the articles cited in this paper presented their supplement to real-world applications. In the digital world, most of people depend on social media to get their valuable data and analyze reviews from these blogs for decision-making.
In the future, from the above survey, a novel hybrid framework that can analyze multiple challenges at one instant could be done. From the above personality prediction survey, authors identified gender and personality prediction based on social media activities. In the future, one can also predict the number of people prone to suicidal attempts and, also, illegal and anti-social activities.

References

[1] Hoang, et.al H.,"*Sarcasm Detection Using Soft Attention-Based Bidirectional Long Short-Term Memory Model with Convolution Network*," IEEE Access,Vol 7, (2019), pp. 23319-23328.

[2] Rajesh, et.al R.,"*Online Public Shaming on Twitter: Detection,Analysis and Mitigation*," IEEE Transactions On Computational Social Systems, (2019) , pp.1-12.

[3] Bouazizi, et.al B., "*A pattern based approach for sarcasm detection in Twitter*," IEEE Access, IEEE Access, Vol 4, (2016), pp.5477-5488.

[4] Fazil, et.al F., "*A Hybrid Approach for Detecting Automated spammer in Twitter*," IEEETransactions on Information Forensics and Security, (2018), pp.1-13.

[5] De-La-Pena-Sordo, et.al D.,"*Anomaly-based user comments detection in social news websites using troll user comments as normality representation*," Oxford University Press Vol. 24, (2016), pp.883-898.

[6] Patxi, et.al P., "*Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying*", Oxford University Press,Vol. 24, (2015), pp. 42-53.

[7] Cynthia Van, et.al C., "*Automatic detection of cyberbullying in social media text*",PLosONE, (2018), pp. 1-22.

[8] Debanjan, et.al D., "*Detecting Personal Intake of medicine from Twitter*," IEEE Intelligent Systems ,Vol. 33, No. 4, (2018), pp. 87 - 95.

[9] Han, et.al H.,"*Deep Learning Model for Classifying Drug Abuse Risk Behavior in Tweets*," IEEE International Conference on Healthcare Informatics (ICHI), New york. USA (2018).

[10] Nhathai, et.al P., "*Enabling Real-Time Drug Abuse Detection in Tweets*," IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA (2017).

[11] Gauri, et.al J., "*Spam detection in social media using convolutionaland long short term memory neural network*," Annals of Mathematics and Artificial Intelligence , Vol. 85, No. 1, (2019), pp. 21-44.

[12] Sreekanth, et.al M.,"*A Neural Network-Based Ensemble Approach for Spam Detection in Twitter*," IEEE Transactions on Computational Social Systems, Vol. 5, (2018), pp. 973-984.

[13] Inuwa-Dutse, et.al I., "*Detection of spam-posting accounts on Twitter*," Neurocomputing, (2018), pp. 1-38.

[14] Michael, et.al M., "*Personality Predictions Based on User Behaviour on the Facebook Social Media Platform*," IEEE Access , Vol. 6, (2018), pp. 61959 - 61969.

[15] Tyshchuk , et.al Y.,"*Modeling Human Behavior on Social Media in Response to Significant Events*," IEEE Transactions on Computational Social Systems,Vol. 5, No. 2, (2018), pp. 444-457.

[16] Dongxu, et.al D.," *Gender Identification via Reposting Behaviors in Social Media*," IEEE Access,Vol. 6 , (2017), pp. 2879-2888.

[17] Young, M., "*The Technical Writer's Handbook*," Mill Valley, CA: University Science (1989).

Follow This Article at The Following Site:

Subbaraju P, Chandra Sekhar K, Raju P, Satyanarayana Raju K, Varma M K S. An Empirical Review on Challenges and Approaches in Sentiment Analysis. IJIEPR. 2020; 31 (2) :317-322
URL: http://ijiepr.iust.ac.ir/article-1-1062-en.html