Sentiment Analysis for Twitter Data

Deep Kaneria, Brijesh Patel

Abstract: With the advancements in web technology and its growth, there's an incredible volume of information present everywhere on the net for internet users and plenty more data is generated on a daily basis. Internet emerged as place for exchanging ideas, sharing opinions, online learning and political views. Social networking sites such as Facebook, Twitter, are rapidly growing as the users are allowed to post and revel their views on various topics, and can discussion with different groups and communities, or post messages across the world. In the area of sentiment analysis large numbers of researchers are working. The main focus is on twitter data for sentiment analysis, that's helpful to research the info within the tweets where opinions are heterogeneous, highly unstructured and are either positive or negative, or neutral in many cases. In this paper, we provide a study and comparative analysis of existing techniques used for opinion mining through machine learning approach. Naive Bayes & Support Vector Machine, we provide research on twitter data.

Keywords: Machine Learning (ML), Naïve Bayes (NB), Twitter, Support Vector Machine (SVM), Sentiment Analysis (SA).

I. INTRODUCTION

Twitter has been very popular and has grown rapidly. An increasing number of people are willing to post their opinions on Twitter, as per the current report [1] 336 million active users monthly, 100 million active users on daily basis and 500 million tweets a day that are considered as a valuable online source for opinions. But the challenging task is extracting and analysing the meaningful insights from Twitter. The natural language used to write these contents and the unstructured nature of the content adds up to the complexity more and it opens new areas for researches like Opinion Mining, Sentiment Analysis etc. In important areas like the prediction of democratic several events, stock exchange, consumer brands, popularity of celebrities, movie box-office, etc. Twitter is a purpose of attraction for many researchers [1]. "Identifying and categorizing opinions computationally, expressed by a portion of text, so as to see the writer's attitude for a specific topic, product, etc. is positive, negative, or neutral is defined by Sentiment Analysis" [2]. Sentiment analysis is applied at different levels starting from a coarse level to a fine level. The coarse level sentiment analysis controls the sentiment of the entire manuscript or document. The fine level sentiment analysis concentrates on the attributes. Sentiment analysis is done on sentence-level that resides in between coarse level and fine level. In the sentiment analysis process, the emotions present within the text can be of two types: Direct and Comparative. The direct sentiments in text from are independent from others within the same sentence [3]. For example, "that's a great smartphone" However, the comparative sentiments in the text denote the comparison of multiple objects in the same sentence. For example, "Bikes are cheaper than the car."

II. RELATED WORK

In current scenario a great amount of research has been done in the domain of sentiment analysis. Many studies show that social media users may purposefully manage their online identity to "look better" than in real life [4], [5]. Other studies show that there is a lack of awareness about managing online identity among college students [6], and that young people usually regard social media as their personal space to hang out with peers outside the sight of parents and teachers [7]. Students' online conversations reveal aspects of their experiences that are not easily seen in formal classroom settings, thus are usually not documented in educational literature. In [8], authors have introduced a hybrid method that is a combination of the usage of sentiment lexicons with a machine learning classifier for polarity detection of subjective texts in the consumer-products domain. In [9], authors have proposed a batch of machine learning methods with semantic analysis to classify the sentence and reviews of different products based on twitter data using WordNet for better accuracy. In [10], authors have proposed an automatic sentiment classifier to classify reviews of Brazilian TV shows into positive or negative category and possessed 90% of accuracy

III. SENTIMENT ANALYSIS

A process that automates mining of opinions, attitudes, emotions, and views from text, speech, tweets, and databases can be called Sentiment Analysis. It includes classifying opinions into categories like "positive, negative or neutral". There are three levels of sentiment analysis: Document-level analysis, Sentence level analysis, and Aspect level analysis. Twitter sentiment analysis falls under the category of sentence-level analysis. The steps for sentimental analysis for twitter data are as below:

Revised Manuscript Received on May 20, 2020.

Deep Kaneria, Department of Computer Engineering, G H Patel College of Engineering & Technology, Vallabh Vidyanagar, Anand, Gujarat, India.

Brijesh Patel, Assistant Professor, Department of Computer Engineering, GCET, Greater Noida, Uttar Pradesh, India.

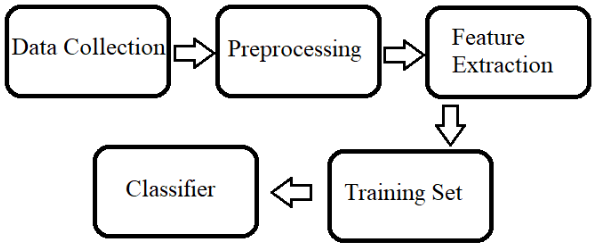


Figure 1. Sentiment Analysis Process

IV. DATA SOURCE

The source of data plays an vital role in sentimental analysis. The three main ways for collection of twitter data are as follow: [11]

- APIs: Two types of APIs are provided by twitter-like Search API and Stream API. Stream API streams real-time data from twitter while the Search API is used for the collection of data using hashtags.
- Data repositories like SNAP, UCI, and Kdnuggets.
- Automated tools such as SocialMention, Tagboard, Radian6, Simplyfy360, and KeyHole.

V. DATA PRE-PROCESSING

The data collected through the above-mentioned tools are always raw, because of that mining is a difficult task. It is important to pre-process or clean the data before applying the classifier. The data cleaning task consist of discarding of hashtags and other notations, stop words, emoticons, URLs, compression of elongated words & decompression of slang words. The pre-processing procedure is as follows.

- Remove notations like hashtags (#), account Id (@) and retweets (RT).
- It's important to remove non-letter data like URLs, hyperlinks, and emoticons as only text data are needed.
- Remove Non-English Tweets.
- Stop words are removed from the datasets, that makes the size of dataset smaller and they play no role in expressing the emotions. Examples are, is, am, etc.
- Compress the stretched-out words such as Yeaahhhh into Yeah.
- Extreme level of sentiments can extracted from slang words as they are nouns or adjectives. Example g8, f9, etc.

VI. FEATURE EXTRACTION

The pre-processed dataset has distinct discrete properties. In feature extraction methods, extraction of different aspects like adjectives, nouns, verbs and later these aspects are identified as negative, neutral or positive to detect the polarity of the sentence. Followings are some of the widely used Feature Extraction methods.

- Parts of Speech (POS): Finding nouns, verbs, adjectives, etc. as they contain a significant amount of sentiments.
- Negative Phrases: Presence of negative words can change the orientation or meaning of sentiment. So, it is necessary to consider them.
- Term Presence and Term Frequency: These features denote individual and distinct words and their occurrences count.
- Words and their Frequencies: Frequency counts of unigrams, bigrams, n-gram models are considered as features. A lot of research is being done on using word presence, instead of frequencies for better describe this feature [12]. showed better results by using presence instead of frequencies.

VII. ALGORITHM

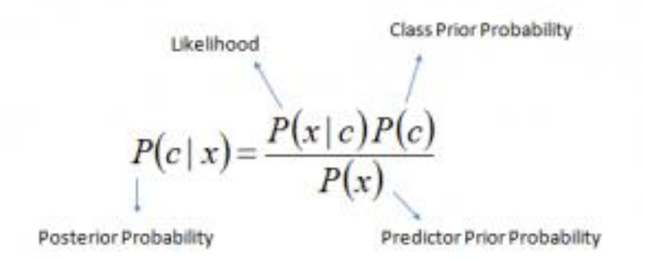
A. Training:

Supervised learning is an important technique for solving classification problems. Training the classifier makes it easier for future predictions for unknown data [13].

B. Naïve Bayes (NB):

It is a classification technique, supported by Bayes' Theorem with an assumption of independence among its predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a specific feature of a class is not related to the presence of the other feature. For example, a fruit can only be said an orange if it's about 3 inches in diameter, round in shape and its orange in colour. If these features are related with each other or on the presence of the opposite features, all the properties sufficiently contribute to the chances that this fruit is an orange and that is why it is known as 'Naive'.

Naive Bayes model is straightforward to create and particularly useful for very large data sets. Along with simplicity, Naive Bayes well known to outperform even highly sophisticated classification methods. Bayes theorem provides the way of calculating posterior probability P(c|x) from P(x), P(c) and P(x|c). Look at the equation below:



$$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Above,

- P(c|x) is posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is a prior probability of class.
- P(x|c) is possibility which is the probability of predictor given class.
- P(x) is a former probability of predictor. [14]

C. Support Vector Machine

Support vector machine (SVM) solves the standard text categorization problem effectively; generally outperforming Naïve Bayes because it supports the concept of maximum margin. The key principle of SVM is to work out a linear separator that splits different classes within the search space with most distance i.e. with maximum margin [8].

If we represent the tweet using t, the hyperplane using h, and classes using a set Cj € {l, -1} into which the tweet has to be classified, the solution is written as follows comparable to the sentiment of the tweet.

$$\vec{h} = \sum i\, ai\, Ci\, \vec{tj}, \qquad ai \geq 0$$

The idea of SVM is to work out a boundary or boundaries that separate distinct clusters or groups of information. SVM performs this task by constructing a group of points and separating those points using mathematical formulas. Fig. 2 illustrates SVM.

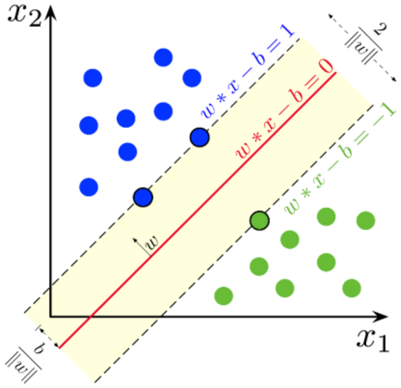


Figure 2. Support Vector Machine [15]

VIII. EVALUATION OF ALGORITHMS

From the study of above-supervised machine learning algorithms used to perform sentimental analysis, we have selected few parameters such as understanding complexity, theoretical accuracy, theoretical training speed, performance with a few numbers of observations and type of the classifier.

Understanding complexity means the technical difficulties to know the algorithm. Theoretical accuracy is that the theoretical measure of how precisely the algorithm can classify on the test set data consistent with the provided training data. Theoretical training speed denotes how rapidly the model will be trained. Performance is related to the precision of the algorithm. In general, accurate algorithms have good performance. Classifier refers to the type of classifier the algorithm belongs to like the linear classifiers, probabilistic classifiers, decision-based classifier [16] [17].

Table 1. Parametric Comparisons of Supervised Machine Learning Algorithms.

Algorithm	NB	SVM
Understanding Complexity	Very Less	High
Theoretical accuracy	Low	High
Theoretical Training Speed	High	High
Performance with small no. of Observations	High	Low

Parametric comparison shown in the Table 1, we can settle that Naïve Bayes algorithms are the modest and easiest to understand and implement, when compared to Support Vector Machine. However, it suffers from lower accuracy because of its simple Bayesian likelihood hypothesis. Whereas Support Vector Machine provides a higher accuracy but it does not provision the automated learning of features [18].

Though, the implementation accuracy of these algorithms extremely depends,on the various factors like domain chosen, data source, size of dataset and pre-processing method applied on the dataset.

IX. DISCUSSION

Sentiment analysis features an extensive variety of applications as following: summarizing review, classifying reviews and other real-time applications. A large amount of applications is possible which may not have been mentioned here. It is found that, sentiment classification mostly depends on domains. From the work, it is obvious that any of the classification models constantly outperforms the other; there are distinct distributions in different types of features. It’s also found that different types of classification algorithms and features can be put together in an effective way to overcome their individual downsides and benefit from other's merits, and most probably enhance the classification performance.

X. FUTURE WORK

In future more, work is required to further improve the measures of performance. There are new applications where sentimental analysis can be applied. The techniques and algorithms used for sentiment analysis are advancing fast. But still few problems of this field remain unsolved. The major challenges are faced when using other languages, handling of negation expressions; produce a summary of opinions relevant to product features/attributes, complexity of sentence, document, handling of implicit product features, etc. More future research might be dedicated to those challenges.

REFERENCES

1. Mitali Desai, Mayuri Mehta, "Techniques of Sentiment Analysis for Twitter Data-A Comprehensive Survey", IEEE on Computing, Communication and Automation, pp.149154, April 2016.
2. Jatinder Kaur, "A Review paper on Twitter Sentiment Analysis,Techniques", International Journal for Research in Applied Science & Engineering Technology, vol.4, pp.137-141, October-2016.
3. C. Romero and S. Ventura, "Educational Data Mining: A Review of the State-of-the-Art," in Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, 2010, vol. 40, no.6, pp. 601-618
4. J.M. DiMicco and D.R. Millen, “Identity Management: Multiple Presentations of Self in Facebook,” Proc. the Int’l ACM Conf. Supporting Group Work, pp. 383-386
5. M. Vorvoreanu and Q. Clark, “Managing Identity Across Social Networks,” Proc. Poster Session at the ACM Conf. Computer Supported Cooperative Work
6. M. Vorvoreanu, Q.M. Clark, and G.A. Boisvenue, “Online Identity Management Literacy for Engineering and Technology Students,” J. Online Eng. Education, vol. 3, article 1
7. M. Ito, H. Horst, M. Bittanti, D. boyd, B. Herr-Stephenson, P.G. Lange, S. Baumer, R. Cody, D. Mahendran, K. Martinez, D. Perkel, C. Sims, and L. Tripp, Living and Learning with New Media: Summary of Findings from the Digital Youth Project. The John D. and Catherine T. MacAuthur Foundation
8. S. Bahrainian and A. Dangel, “Sentiment Analysis using Sentiment Features”, in Int. joint Conf. of Web Intelligence and Intelligent Agent Technologies, 2013, pp. 26-29.
9. G. Gautam and D. Yadav, “Sentiment analysis of twitter data using machine learning approaches and semantic analysis”, in 7th Int. Conf. on Contemporary Computing, 2014, pp. 437-442.
10. A.C.E.S Lima. and L.N. de Castro, “Automatic sentiment analysis of Twitter messages”, in 4th Int. Conf. on Computational Aspects of Social Networks (CASoN), 2012
11. “Three Cool and Inexpensive Tools to Track Twitter Hashtags”, June 11, 2019. [Online]. Available

http://dannybrown.me/2013/06/11/threecool-toolstwitterhashtags/ [Accessed: 3-Dec-2019].

12. Pang, B. and Lee, L. “A sentimental education: Sentiment analysis using subjectivity,summarization based on minimum cuts”. 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04), pp. 271-278, 2004,
13. Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of twitter data: a survey of techniques." arXiv preprint arXiv:1601.06971 (2016).
14. Ray, Sunil, “6 Easy-Steps to Learn Naive Bayes Algorithm (with Code in Python).” Analytics Vidhya, 11 Apr. 2018, www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.
15. Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks", Machine Learning, 20 (3): pp. 273–297.
16. N. Kasture and P. Bhilare, “An Approach for Sentiment analysis on a social networking site”, Computing Communication Control and Automation (ICCUBEA), pp. 390-395, 2015.
17. X. Chen, M. Vorvoreanu and K. Madhavan, “Mining Social Media Data to Understand Students’ Learning Experiences”, IEEE Transaction, vol. 7, no. 3, pp. 246-259, 2014.
18. V. Singh and S. K. Dubey, “Opinion mining and analysis: A literature review”, in 5th Int. Conf. on Confluence The Next Generation Information,Technology Summit (Confluence), pp. 232-239, 2014

AUTHORS PROFILE

Deep Kaneria is the student pursuing Maters of Computer Engineering from the G H Patel college of Engineering & Technology. His research work and interest mainly focused on Machine Learning, Image Processing, Computer Vision and Natural Language Processing

Brijesh Patel received the Bachelor degree from the Department of Computer Engineering, Veer Narmad South Gujarat University, Surat, India, in 2008, and the Master degree in computer engineering from the Gujarat Technological University (GTU), Ahmedabad, in 2012. He is pursuing a PhD from GTU. He is currently an Assistant Professor with the department of computer engineering, GCET. His research interests include location management in computer network, network security, network traffic classification, machine learning, deep packet inspection.