A Review on Sentiment Analysis of Twitter Dataset Based on Classification Technique

Sonu Patidar#1, Sanjay Bansal*2
#*Department of Computer Science, RGPV University
Acropolis Institute of Technology and Research, Indore, India
1 15shona.patidar@gmail.com
2 sanjaybansal@acropolis.in

Abstract **-Social networks have revolutionized the way in which people communicate. Information available from social networks is beneficial for analysis of user opinion, looking at the response to policy change or the enjoyment of an ongoing event. Manually sifting through this data is tedious and potentially expensive. Sentiment analysis is a relatively new area, which deals with extracting user opinion automatically. Emotion sensing or sentiment analysis is a complicated task of machine learning technology. That belongs to the Natural language processing branch of artificial intelligence. It is a broad domain of learning and analysis. Among the text classification and their emotional orientation discovery is also part of this domain. In this paper, we proposed decision tree algorithm based text classification model for performing sentiment on twitter based data. Additionally the comparative performance is also measured with the traditional ID3 algorithm and similar variant of the improved ID3 classification algorithm. In order to compare the performance of the algorithms the accuracy, error rate, memory consumption and time consumption is taken as stand parameters.**

Keywords: **ID3, Sentiment, Tweeter, Social Media, Data Mining, Text Mining, NLP.**

I. INTRODUCTION

TWITTER [1] is a social networking application which allows people to micro-blog about a broad range of topics. Micro-blogging is defined as "a form of blogging that lets you write brief text updates about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web. Twitter helps users to connect with other Twitter users around the globe. Successful micro-blogging services such as Twitter have become an integral part of the daily life of millions of users. In addition to communicating with friends, family or acquaintances, micro-blogging services are used as recommendation services, real-time news sources and con tent sharing venues.

These tweets tend to spread to a large number of users in very little time. Users on Twitter not only tweet about their personal issues or nearby events, but also about more general topics or news [2]. Due to the large amounts and diversity of real-time information contained on the site, Twitter lists a freshly updated set of trending topics. Trending topics comprise the top terms being discussed currently on Twitter. This list of top terms, which is updated in real-time, provides a reflection of the current main interests of the community, i.e., the most- discussed conversations right at the moment.

A General model for sentiment analysis is as follows:

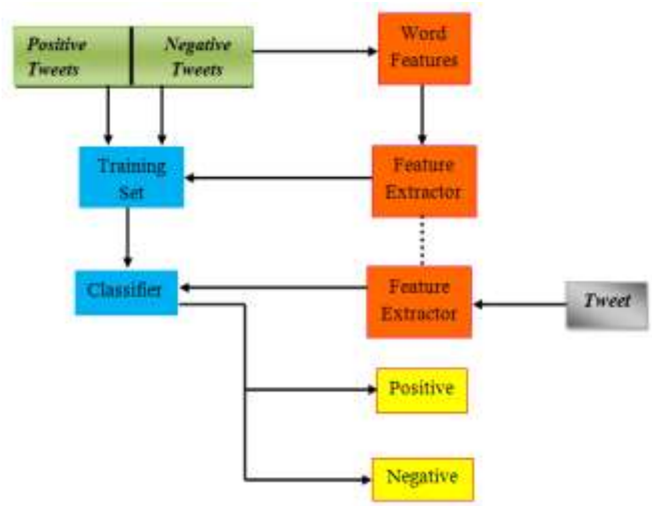

Figure 2.2Sentiment Analysis Architecture [23]

Following are the phases required for sentiment analysis of twitter data:

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points

- ✓ Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
- ✓ Correct the spellings; sequence of repeated characters is to be handled Replace all the emoticons with their sentiment.
- ✓ Remove all punctuations, symbols, and numbers
- ✓ Remove Stop Words

- ✓ Expand Acronyms (we can use a acronym dictionary)
- ✓ Remove Non-English Tweets

A. *Characteristics of Tweets*

Twitter messages have many unique attributes, which differentiates our research from previous research [10]:

Length The maximum length of a Twitter message is 140 characters. From our training set, we calculate that the average length of a tweet is 14 words or 78 characters. This is very different from the previous sentiment classification research that focused on classifying longer bodies of work, such as movie reviews.

Data availability another difference is the magnitude of data available. With the Twitter API, it is very easy to collect millions of tweets for training. In past research, tests only consisted of thousands of training items.

Language model Twitter users post messages from many different media, including their cell phones. The frequency of misspellings and slang in tweets is much higher than in other domains.

Domain Twitter users post short messages about a variety of topics unlike other sites which are tailored to a specific topic. This differs from a large percentage of past research, which focused on specific domains such as movie reviews.

The use of digital text is increases as the social media increases their effect in daily life. A number of research groups and individual researchers are working to finding the patterns on these data. In this study the social media text analysis and sentiment analysis techniques are investigated and a new classification technique is proposed for enhancing the performance of text classification. The given chapter provides an overview of the proposed work and involved investigation.

B. *Sentiment Analysis*

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction. The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

- ✓ *Opinion:* A conclusion open to dispute (because different experts have different opinions)
- ✓ *View:* subjective opinion
- ✓ *Belief:* deliberate acceptance and intellectual assent
- ✓ *Sentiment:* opinion representing one's feelings

Sentiment Analysis is a term that includes many tasks such as sentiment extraction, sentiment classification, and subjectivity classification, summarization of opinions or opinion spam detection, among others. It aims to analyze people's sentiments, attitudes, opinions emotions, etc. towards elements such as, products, individuals, topics, organizations, and services [3].

C. *Feature Extraction*

The preprocessed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram [4]. Machine learning techniques require representing the key features of text or documents for processing. These key features are c o n s i d e r e d as feature vectors which are used for the classification task. Panget al. [5] showed better results by using presence instead of frequencies.

- ✓ ***Parts of Speech Tags***: Parts of speech like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.
- ✓ ***Opinion Words and Phrases:*** Apart from specific words, some phrases and idioms which convey sentiments can be used as features. E.g. cost someone an arm and leg.
- ✓ ***Position of Terms:*** The position of a term with in a text can effect on how much the term makes difference in overall sentiment of the text.
- ✓ ***Negation:*** Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion. e.g., I am not happy
- ✓ ***Syntax****:* Syntactic patterns like collocations are used as features to learn subjectivity patterns by many of the researchers.

II. LITERATURE SURVEY

Agarwal et al. [7] present an analysis in which twitter is different to other forms of raw data which are used for sentiment analysis as sentiments are conveyed in one or two sentence blurbs rather than paragraphs. Twitter is much more informal and less consistent in terms of language. Users cover a wide array of topics which interest them and use many symbols such as emoticons to express their views on many aspects of their life When using human generated status updates, sentiment are not always obvious; many tweets are ambiguous and can use humors to maximize the opinion to other human readers but deflect the opinion to a machine learning algorithm.

ApoorvAgarwal et al [8] examine sentiment analysis on Twitter data. The contributions of this paper are: (1) first introduce POS-specific prior polarity features. (2) And explore the use of a tree kernel to obviate the need for tedious feature engineering. The new features (in conjunction with previously proposed features) and the tree kernel perform approximately at the same level, both outperforming the state-of-the-art baseline.

The influence of micro blog on information transmission is becoming more and more obvious. By characterizing the behavior of following and being followed as out-degree and in-degree respectively, a micro blog social network was built in this paper. It was found to have short diameter of connected graph, short average path length and high average clustering coefficient. The distributions of out-degree, in-degree and total number of micro blogs posted present power-law characters. The exponent of total number distribution of micro blogs is negatively correlated with the degree of each user. With the increase of degree, the exponent decreases much slower. Based on empirical analysis, Qiang Yan et al [9]proposed a social network based human dynamics model in this paper, and pointed out that inducing drive and spontaneous drive lead to the behavior of posting micro blogs. The simulation results of model match well with practical situation.

Another consideration when using a dataset generated from Twitter is that a considerably large amount of tweets which convey no sentiment such as linking to a news article, which can lead to difficulties in data gathering, training and testing. Movassate et al. [10] provides Sentiment analysis of tracking opinions and attitudes on the web and determines if they are positively or negatively received by the public.

III. PROPOSED WORK

Engineered by Ross Quinlan the ID3 is a straightforward decision tree learning algorithm. The main concept of this algorithm is construction of the decision tree through implementing a top-down, greedy search by the provided sets for testing every attribute at each node of decision. With the aim of selecting the attribute which is most useful to classify a provided set of data, a metric is introduced named as Information Gain [38].

To acquire the finest way for classification of learning set, one requires to act for minimizing the fired question (i.e. to minimize depth of the tree). Hence, some functions are needed that is capable of determine which questions will offer the generally unbiased splitting. One such function is information gain metric.

ID3 Decision Tree

Input: Examples, Target Attribute, Attributes

Output: Decision Tree

Process:

- Produce a node being root node of the tree
- Check if all the examples are positive, If yes then generate a single node tree, ROOT, having label = +.
- In case all the examples are found negative, then make a single node tree, ROOT, with label = -.
- If there are no attributes for prediction, then create a single node tree ROOT labeled as the most ordinary used value for that attribute.
- Else Start following procedure
    - M = an attribute that is classifying the Examples in Best way.
    - Make M the decision tree attribute
    - Repeat for every probable value, v_i, of M,
    - Expand the Root with one branch, equivalent to the test M = v_i.
    - Let Examples(v_i) be a subset of examples that have the value v_i for M
    - If Examples(v_i) is empty
    - Add a leaf node under the new branch, labeled with most ordinary value of the attribute in the examples
    - Otherwise add the sub-tree ID3 (Examples(v_i), Target Attribute, Attributes – {A}), under this new branch
- End
- Return Root

IV. CONCLUSION

Data mining offers the supervised and unsupervised learning concept to analyses the data and classifies or categorize in a predefined groups of data. The algorithms enable us to use the computer based algorithms to analyze the data automatically without any human efforts. The proposed work analyzes the social network based text for their sentiments and orientation based text classification. Therefore the proposed work involves the pre-processing, tagging, learning and the classification of newly arrived patterns. The performance of the system is estimated for finding the system accuracy and error rate for the sentiment data.

ACKNOWLEDGEMENT

The authors wish to thank Prof. Abhay Kothari for sharing his ideas in writing this paper.

REFERENCES

[1] http://www.twitter.com
[2] E. Mishaud. Twitter: Expressions of the whole self, Master's thesis, Department of Media and Communications, University of London, 2007.
[3] Vishal A. Kharde and S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (IJCA) Volume 139 – No.11, April 2016
[4] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment Treebank" Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2013
[5] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Aug 23-24 2014, pp 171-175.
[6] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.
[7] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data," Proceedings of the workshop on languages in social media, Association for Computational Linguistics, 2011.
[8] Adam Marcus, Michael S. Bernstein, Osama Badar,David R. Karger, Samuel Madden, Robert C. Miller, "TwitInfo: Aggregating and Visualizing Microblogsfor Event Exploration", CHI 2011, May 7–12, 2011, Vancouver, BC, Canada. ACM
[9] Junghoon Chae, Dennis Thom, Yun Jang, SungYe Kim, Thomas Ertl, David S. Ebert, "Public behavior response analysis in disaster events utilizing visual analytics of microblog data", & 2013.
[10] Parikh, Ravi, and Matin Movassate. "Sentiment analysis of user-generated twitter updates using various classification techniques." CS224N Final Report (2009): 1-18.
[11] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, 2005
[12] Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009
[13] P. Bhargavi, B. Jyothi, S. Jyothi, K. Sekar, "Knowledge Extraction Using Rule Based Decision Tree Approach", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2000
[14] Umajancy. S, Dr. Antony SelvadossThanamani, "An Analysis on Text Mining –Text Retrieval and Text Extraction", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013
[15] H. Karanikas, C. Tjortjis, and B. Theodoulidis, "An approach to text mining using information extraction," in Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference, 2000.
[16] R. Hale, "Text mining: Getting more value from literature resources," Drug Discovery Today, vol. 10, no. 6, pp. 377–379, 2005.
[17] S. Chakrabarti, Mining the Web: Analysis of Hypertext and Semi Structured Data. San Francisco, CA: Morgan Kaufman, 2000.

