Review Paper on Sentiment Analysis of Twitter Data Using Text Mining and Hybrid Classification Approach

Shubham Goyal
Department of Computer Science, ASRA College of Engineering and Technology, Bhwanigarh, Punjab, India

Abstract – In Sentiment analysis we use natural language processing and information to extracting writer's comments or reviews. In this paper we use Data text mining and hybrid approach of KNN Algorithm and Naïve Bayes Algorithm to find the sentiments of Indian people on Tweeter.

Keywords: Sentiment Analysis, Text Mining.

1. INTRODUCTION

Human life is filled with emotions and opinions. People love to share their emotions and opinions at every place but social media is one of the most common and easy way to share our feelings. Today people not only comment on the existing information, bookmark pages and provide ratings but they also share their ideas, news and knowledge with the community at large. In this way, the entire community becomes a writer, in addition to being a reader [1].

The existing mediums like Blogs, Wikis, Forums and Social Networks where users can post information, give opinions and get feedback from other users on different topics, ranging from politics and health to product reviews and travelling. Recently, many researchers have focused on this area [2]. They are trying to fetch opinion information to analyze and summarize the opinions expressed automatically with computers. This new research domain is usually called Opinion Mining and Sentiment Analysis [4].

Until now, researchers have evolved several techniques to the solution of the problem. Current-day Opinion Mining and Sentiment Analysis is a field of study at the crossroad of Information Retrieval (IR) and Natural Language Processing (NLP).

2. LITERATURE SURVEY

Ortigosa and Alvaro et. al [2] proposed a novel method for sentiment analysis in social site giant Facebook that, starting from the messages written by its users, supports: (i) to extract useful information about the Facebook users' sentiment polarity (whether it is positive, neutral or negative), which reflected from the messages written by users; and (ii) to model the users' normal sentiment polarity and to analyze significant emotional changes in user.

Pak and Alexander et al. proposed [3] By using the corpus, Author builds a sentiment classifier, which is capable of determining positive, neutral and negative sentiments for the whole document. Experimental results show that the proposed techniques are more efficient and perform better as compared to previously proposed techniques.

Agarwal and Apoorv [5] explained one such popular micro blog named as Twitter

and build models to classifying the "tweets" into positive and negative sentiment or they can be neutral. Author build novel models for two classification: first one is a binary task of classifying sentiment of users into positive and negative classes and second is a 3-way task of classifying sentiment of users into positive, negative and neutral. Author experiment with two types of models: (1) unigram model which is a feature based model (2) a tree 30 kernel based model.

Aisopos and Fotis et al. [6] presented with Microblog content, some serious challenges are associated. Some of these are the applicability of sentiment analysis used in past and different classification methods caused by their inherent characteristics of content. To resolve them, author introduces a method that relies on two orthogonal and complementary sources of evidence: context-based method captured by polarity ratio and content-based features acquired by n-gram graphs. Both the methods are language-neutral and tolerant to noise; guarantee high robustness and effectiveness in the manner author are considering.

Horakova and Marketa et al. [7] Present a model which collects tweets from social networking sites and thus provide a view of business intelligence. In our framework, there are two layers in the sentiment analysis tool, the data processing layer and sentiment analysis layer. Data processing layer deals with data collection and data mining.

3. PROPOSED METHODOLOGY

The first step starts with the extraction of tweets followed by preprocessing of the extracted tweets. Then Classifier algorithm has to be applied on it.

KNN Algorithm

KNN is type of instance based learning or lazy learning. In this learning the function is approximately locally and all computation is deferred until classification. It is simplest of all machine learning algorithms. In KNN classification, the output is class membership. An object is classified by majority votes of its neighbors by the object being assigned to class most common among its k nearest neighbor (k is positive small integer). The nearest neighbor is determined using similarity measure usually distance functions are user. Following are the distance function used by KNN.

Euclidean distance function

$$\sqrt{\sum_{i=1}^{N}(a_i - b_i)^2} \quad (1)$$

Manhattan distance function

$$\sum_{i=}^{N}|a_i - b_i| \quad (2)$$

Where $\{(a_1, b_1), (a_2,b_2), (a_3,b_3) \ldots\ldots (a_N, b_N)\}$ is training datasets.

Naïve Bayes Algorithm

The algorithm is named after famous statistician Thomas Bayes who proposed Bayesian theorem. This theorem assumes that all the attributes are conditionally independent to each other. In this algorithm, conditional probability for each

attribute with respect to certain class level is calculated. The new document is classified using sum of probabilities for each class. The classification framework is briefly discussed as follows:

Suppose we have D set of tuples and each tuple has attribute vector X(x1, x2, x3 , ... xn) of n dimensions. Let there are k number of classes C1, C2, C3... Ck. The classifier predicts X belongs to Ci if

$$P\left(\frac{C_i}{X}\right) = P\left(\frac{C_j}{X}\right) \text{ for } 1<=j<=k, j<>I \quad (3)$$

Posterior probability is calculated as

$$P\left(\frac{C_i}{X}\right) = \frac{P(X/Ci)\ P(c_i)}{P(X)} \quad (4)$$

4. CONCLUSION

In this survey we found that social media like twitter can be used to predict the sentiments of people. We will use combination of KNN ALGORITHM and NAÏVE BAYES ALGORTHM to find the result. So our proposed system concludes the sentiments of tweets which are extracted from twitter using its API. We also implemented features like to find emotions, smileys; injections as they are recently become a huge part of internet.

5. REFERENCES

1. Scholar, P. G. "Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data."
2. Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." *Computers in Human Behavior* 31 (2014): 527-541.
3. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREC*. Vol. 10. 2010.
4. Jebaseeli, A. Nisha, and E. Kirubakaran. "A Survey on Sentiment Analysis of (Product) Reviews." *International Journal of Computer Applications* 47.11 (2012).
5. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011.
6. Aisopos, Fotis, et al. "Content vs. context for sentiment analysis: a comparative analysis over microblogs." *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 2012.
7. Jebaseeli, A. Nisha, and E. Kirubakaran. "A Survey on Sentiment Analysis of (Product) Reviews." *International Journal of Computer Applications* 47.11 (2012).