# ANALYSIS OF TECHNIQUES IN SENTIMENT VARIATIONS

**Thejas Mol Thomas[1], Pretty Babu[2]**

*[1]Department of Computer Science and Engineering,*
*Sree Buddha College of Engineering, (India)*
*[2]Assistant Professor, Department of Computer Science and Engineering,*
*Sree Buddha College of Engineering (India)*

## ABSTRACT

*Data mining is one of the emerging technology in the area of data mining. Twitter is a popular micro blogging site where people share their views and opinions. The retrieving or filtering of data from twitter is a challenging task since it may contain noise data and also it is difficult to represent the data. Thus the finding of variation in the opinions or views of the people in these micro blogging sites is a complex task. The main aim of this paper is to analyse the techniques used for retrieving and finding the sentiment variation behind the opinions of the user. The techniques used for finding variations are LDA, K-clustering , POS and cosine similarity.*

***Keywords:* Clustering, Cosine Similarity, LDA, Part-Of-Speech Tagging.**

## I. INTRODUCTION

Data mining is a vast area which deals with the mining or finding hidden data in large collection of data. The general method of data mining process is to extract or mine information from a large data set and transform it into a structure for future use. Besides the raw data analysis step, it includes database and data management methods, data pre-processing, modeling and inference computations, interestingness metrics, complexity calculations, post-processing of exposed structures, graphical representation, and online revision. Out of this, text mining is the most emerging technology which is used to mine text data from the web. There are many ways for text analytics such as survey and many research areas and business uses this information for their purpose. Then also this data retrieving has many challenges. Tweets are frequently used to express a public's emotion. It describes a diversity of new sources of online data that are designed, initiated, published and used by customers. Sentiment variation on twitter data has provided an effective way to expose public opinion. It is generally difficult to find the exact causes of sentiment variations since they may involve complicated internal and external factors. We observed that the emerging topics in the variation period is highly related to the legitimate reasons responsible for the variations. When people specify their opinions, they usually mention reasons that support their current views. Mining emerging events/topics is a challenging task: (1) The tweets present in the variation period can be very noisy, which covers irrelevant "background" topics which had been considered for a long time and it does not contribute to the changes of the public's opinion. (2) The events and topics which is related to the opinion variations are hard to represent or model. Keywords produced through topic modelling can depict the underlying events to some extent. But they are not as intuitive as natural language sentences. (3) Reasons could be complicated and involve a number of events. These events might not be equally important. Therefore, the mined events should be ranked with respect to their contributions.

The paper aims at analysing the techniques used in finding the variations in the tweets. For analysing we choose the methods : LDA, K clustering, cosine similarity and POS. Each method has its own benefits. In this analysis, we define document similarity as the distance between topics or words within documents based on the uniformity of their meaning or well formed content. Accordingly, when certain sets of documents display high correlation values, it means that they are semantically identical. We also focus at analysing the combination of these methods can be efficiently used to track the sentiments in tweets which is helpful in finding the variations in the opinion of people.

## II. SYSTEM ANALYSIS

The methods used to reduce the challenges in mining was able to solve the problems up to a certain level. But these methods didn't gave a complete solution and it raised certain problems. The current system raises security issues like only English tweets are taken, considers only marketing tweets, computation of vectors and also discuses only about the properties. The existing system uses a combination of data sets, data extraction using keywords, classification and part-of-speech tagging. Hence we analysis of few methods which can be combined to form a new model to efficiently find the variations of sentiment in the tweets. We use Twitter as the textual data source for our analysis, because it is one of the most popular micro blog worldwide and the topics on which Twitter users post are not limited.

The analysis of the system uses following techniques.

2.1 LDA
2.2 K Clustering
2.3 Cosine Similarity
2.4 Part-of-speech Tagging

These methods are combined together to form a new system which can be implemented to discriminate the properties and variations of public sentiments in twitter.

### 2.1 Latent Dirichlet Allocation (LDA)

LDA is a arable probabilistic model for collecting distinct data with a three-level hierarchical Bayesian model, where each item of a collection is modeled as a definite mixture over an underlying set of topics or words. This technique is often used in the text modeling framework, while the topic probabilities imply an accurate representation of a document. We apply this technique to discover the underlying topics or words in the word sets where people describe their subjective views. The goal of this analysis is to draw apparent representations for both word sets including user opinion with the positive or negative semantic properties.

The model can be represented using plate notation. With plate notation, the dependencies among variables which can be captured precisely. The boxes are "plates" representing duplicates. The outer plate represents documents, while the inner plate represents the imitated choice of topics and words within a document. M denotes the number of documents and N represents the number of words in a document. Thus:

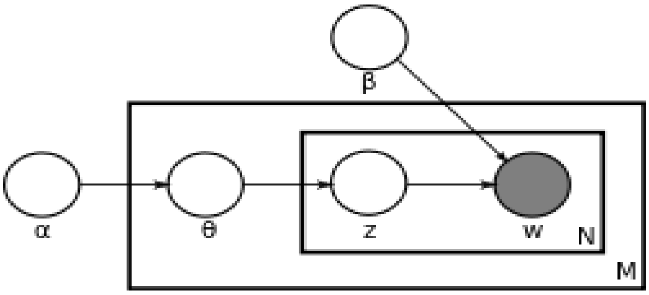


**Fig 1: LDA model**

$\alpha$ is the parameter of the per-document topic distributions,

$\beta$ is the parameter of the per-topic word distribution,

$\theta_i$is the topic distribution for corresponding document *i*,

$\phi_k$is the word distribution for topic *k* in the document,

$z_{ij}$is the topic for the *j*th word in document *i*, and

$w_{ij}$is the specific word in the distribution.

The result using LDA can be shown using a graph. In several researches they have used LDA to retrieve topics and shown the sentiment variations. One of the example graph using LDA is shown below. The graph shows the positive and negative sentiment variations. The result is about 85 percent accurate compared to old text retrieving methods.

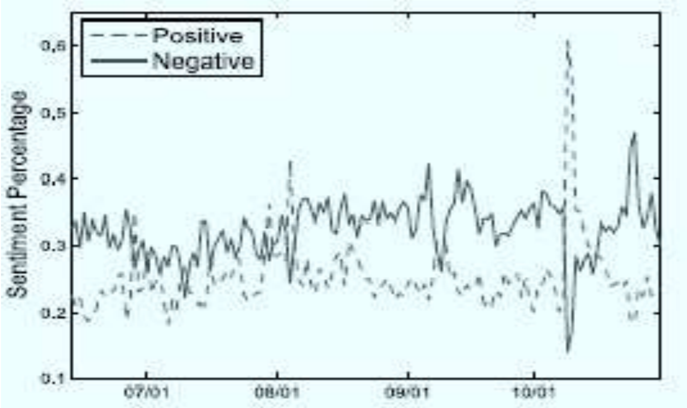


**Fig 2: Graph showing sentiment variation using LDA**

## 2.2 K-Means Clustering

After finding the word sets with the same sentiment polarities look to be more correlated, hence we are further curious to find how word sets with different sentiment signs are segmented. In detail, we would like to understand how accurately two kinds of word sets with different sentiment properties are gathered together. We propose a popular clustering method, k-means. K-means clustering is one of the data mining techniques popularly used to divide a data set into *k* groups in such a way that reduces the within-cluster sum of squares (WCSS):

$$argmin \sum_{i=1}^{k} \sum_{xj} \left\| x_{i-} \mu_j \right\|^2$$

where $\mu i$ is the mean of points in *Si* and (**x**1, **x**2… **x***n*) is a set of observations.

Thus, the k-means method segments the data set based on the frequency of the terms that appear in the document matrix. We set k at two, hypothesizing that there would be two segments and that the word sets originating from the data sets with negative sentiment values would produce one segment, and the word sets from the data sets with positive values would create another segment. Figure illustrates the k-means analysis result that the positive word sets create one cluster, cluster1, and the negative word sets create another cluster, cluster 2, while k-means divides the word sets one to twelve into two groups exactly depending on their sentiment polarities.

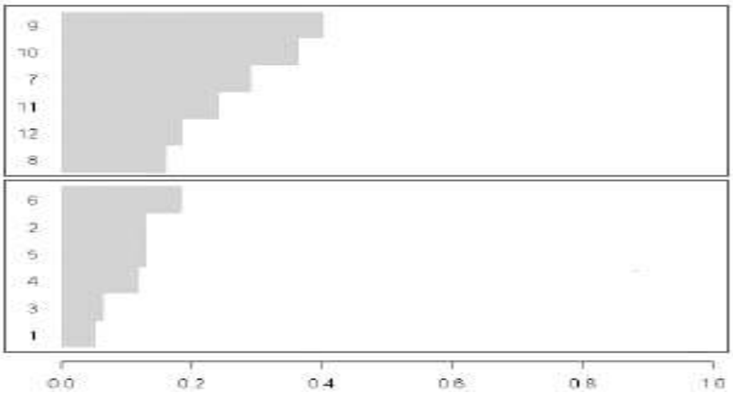
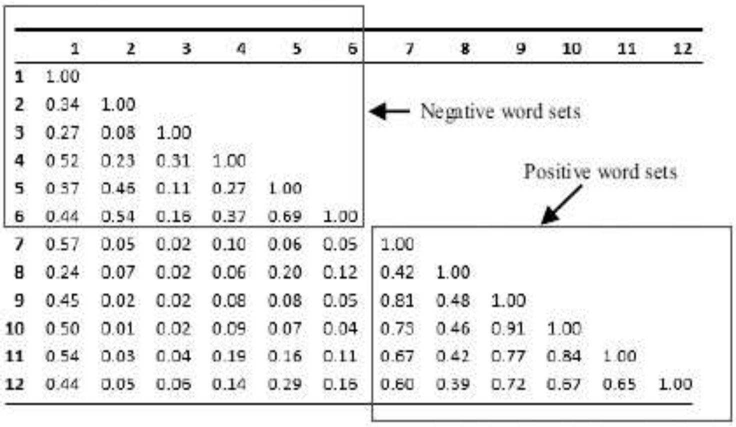


**Fig 3: Graph showing K clustering of word sets**

### 2.3 Cosine Similarity

Cosine similarity is a metric frequently used to discover similarity and dissimilarity textual data. This metric basically calculates the cosine of the angle between two vectors, indicating that cosine 0 degree represents cosines similarity value of 1, which implies that two vectors are exactly the same, and cosine 90 degrees, a cosine similarity value of 0, which means that the vectors are completely independent. Specifically, cosine similarity measures the inner product space between two vectors which are derived from documents. The set of documents is represented as a set of vectors in a vector space where two documents are relatively close in space whenever they are similar in terms of the semantic meaning. For example, vec1= [1,1,1,1,1,2,1,0,0] and vec2 = [1,1,1,2,0,0,1,1,1] have similarity of 0.9487, which is derived from formula . In general, cosine similarity is calculated based on following formula, where *A* and *B* represent two vectors values.

$$similarity = \cos\theta = \frac{A * B}{\|A\| \, \|B\|}$$



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | | | | | | | | | | | |
| 2 | 0.34 | 1.00 | | | | | | | | | | |
| 3 | 0.27 | 0.08 | 1.00 | | | | | | | | | |
| 4 | 0.52 | 0.23 | 0.31 | 1.00 | | | | | | | | |
| 5 | 0.37 | 0.46 | 0.11 | 0.27 | 1.00 | | | | | | | |
| 6 | 0.44 | 0.54 | 0.16 | 0.37 | 0.69 | 1.00 | | | | | | |
| 7 | 0.57 | 0.05 | 0.02 | 0.10 | 0.06 | 0.05 | 1.00 | | | | | |
| 8 | 0.24 | 0.07 | 0.02 | 0.06 | 0.20 | 0.12 | 0.42 | 1.00 | | | | |
| 9 | 0.45 | 0.02 | 0.02 | 0.08 | 0.08 | 0.05 | 0.81 | 0.48 | 1.00 | | | |
| 10 | 0.50 | 0.01 | 0.02 | 0.09 | 0.07 | 0.04 | 0.73 | 0.46 | 0.91 | 1.00 | | |
| 11 | 0.54 | 0.03 | 0.04 | 0.19 | 0.16 | 0.11 | 0.67 | 0.42 | 0.77 | 0.84 | 1.00 | |
| 12 | 0.44 | 0.05 | 0.06 | 0.14 | 0.29 | 0.16 | 0.60 | 0.39 | 0.72 | 0.67 | 0.65 | 1.00 |

Negative word sets
Positive word sets

**Fig 4: Table showing cosine similarity of words [1]**

We apply the cosine similarity method to the term document matrix to determine similarity and dissimilarity between two forms of word sets. The following figure shows the cosine similarity result.

### 2.4 Part-Of-Speech Tagging

Another method used in the data manipulation process is part-of-speech (POS) tagging. POS tagging is one form of syntactic analysis that reads text in some language and assigns parts of speech to each word (or each token) such as noun, verb, adjective, etc. Figure 2 illustrates how POS assigns a tag to each word where "VB" represents verb, "NN" or "NNS" common noun, "JJS" adjective, "IN" preposition, and so forth.

"take/VB", "#valentine/JJ", "Best/JJS",
"Price/NN" "Bvlgari/NNP",
"Women/NNP", "Eau/NNP",
"De/NNP", Perfume/NNP",
"Spray/NNP", "by/IN",
"searched/VBN", "and/CC". "for/IN",
"Guide/NNP", "to/TO"

**Fig 5: Figure showing the POS Tagging**

We apply this technique to produce final data sets consisting of only adjectives, adverbs, and verbs in each document (i.e., a tweet).

## III. CONCLUSIONS

The overall aim of the data mining process is to excerpt information from a large data set and transform it into an logical structure for further uses. The mining is done in twitter data set. The system is used to discriminate the properties and to analyse the public sentiment variations. Thus, in proposed work, we analysed four methods: Latent Dirichlet Allocation (LDA) based models, Cosine similarity, POS and Clustering. The LDA model can filter out background topics and then extract foreground topics to reveal possible reasons. To give a more intuitive representation. Our proposed models were evaluated on real Twitter data. Experimental results showed that our models can mine possible reasons behind sentiment variations. Moreover, the proposed models are general: they can be used to discover special topics or aspects in one text collection in comparison with another background text collection. Also these methods can be combined together to form a new model which can be used for efficient sentiment tracking in twitter as well as in other social networking sites.

## IV. ACKNOWLEDGMENT

We would like to thank various technological experts who researches about data mining and sentiment variation and improve the result by implementing new methods. We would also like to thank Google for providing details on different issues addressing the challenges of mining data from twitter and about other related techniques.

## REFERENCES

[1] Eun Hee Koand Diego Klabjan, "Semantic Properties of Customer Sentiment in Tweets", 28th International Conference on Advanced Information Networking and Applications, 2014.

[2] A.Ghose, and S.P. Han, "An empirical analysis of user content generation and usage behavior on the mobile internet", Management Science, vol. 57, September 2011.

[3] B. Liu, M. Hu, and J. Cheng, Opinion observer: analyzing and comparing opinions on the web, 2005.

[4] B. Liu, "Sentiment analysis and opinion mining" San Rafael, CA: Morgan & Claypool Publishers, 2012.

[5] B.O'Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith, "From tweets to polls: linking text sentiment to public opinion time series" , International AAAI Conference on Weblogs and Soial Media, May 2010.

[6] D.M. Blei, A.Y. Ng., and M.I., Jordan, "Latent dirichlet allocation", The Journal of Machine Learning Research, vol. 3, pp. 993-1022, March 2003.

[7] G. Mishne and N. Glance, Leave a reply: an analysis of weblog comments,WWW'06, 2006.

[8] J.MacQueen, "Some methods of classificatioin and analysis of multivariate observations", L.M.LeCam and J.Neyman, editors, Proc. 5th Berkeley Symposium on Math., Stat., and Prob., p. 281. U. California Press, Berkeley, CA, 1967.

[9] K. Gimpel et al.,"Part-of-speech tagging for Twitter: annotation, features, and experiments", HLT'11Computational Linguistics: Human Language Technologies, vol. 2, pp. 42-47, 2010.

[10] P. Blackshaw, and M. Nazzaro, Consumer- generated media (CGM) 101: word of mouth is the ace of the web-fortified consumer, Intelliseek White Paper 2004.

[11] P.D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews Association for Computational Linguistics, 2002.