PESTLE BASED EVENT DETECTION AND CLASSIFICATION

Vaishali Ugale[1], Soudamini Pawar[2]

[1]*ME student, Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi, Savitribai Phule Pune University, Pune, Maharashtra, India*

[2]*Assistant Professor, D. Y. Patil College of Engineering, Akurdi, Savitribai Phule Pune University, Pune, Maharashtra, India*

Abstract

Organizations use PESTLE classification as a tool for tracking the environment in which they are functioning and for launching plan of new product or service. It helps to give true view of the environment from different aspects. These aspects are essential for any business that organization may be in as it gives a clear picture one wishes to check and observe while contemplating on certain idea or plan. The PESTLE framework helps to understand the market dynamics and is also one of the pillars of strategic management of an enterprise that drives goal and strategy for them. PESTLE based event detection approach proposed in this paper would help for PESTLE analysis of any organization. It puts together all relevant factors in terms of detected events in one place and classifies them into separate buckets while taking current market situation into consideration. We accomplish this with the application of clustering technique and later training the classifier to classify the events in PESTLE format.

Keywords: *Event Detection, PESTLE Analysis, Twitter*

----------***----------

1. INTRODUCTION

Various types of social media sites which include traditional media such as newspaper, radio and television and non-traditional media such as Facebook, Twitter together form the social media. Among these media sites, on-line Social Networks are web-based services that make it possible for individuals and communities to connect with real world friends and acquaintances on-line like Facebook, Linkedin, Orkut etc. The relatively new phenomenon is microblogging that can be considered as a counterpart to blogging, but the contents permitted are limited. However the usage of these social media sites can be from communication medium and social interaction to citizen journalism.

Social media sites (e.g., Twitter, Facebook, and YouTube) have emerged as powerful means of communication in recent years for people who are looking to share and exchange information on a wide variety of real-world events. These events may be popular, widely known ones (e.g., a concert by a popular music band) or of smaller scale local events (e.g., a local social gathering, a protest, or an accident). These events are typically reflected by Twitter with short messages posted on it as they happen. Thus, the content of such social media sites can be particularly useful for real-time identification of events and the messages contributed by associated user. An ever increasing amount of content captured during and associated with various types of events is brought to the web. This is only possible due to the ease of publishing content on social media sites. Event content shared on social media sites vary widely, ranging from planned, known occurrences such as a concert or a parade, to spontaneous, unplanned incidents such as an earthquake. By automatically identifying and characterizing these events, a rich search and presentation of all event content can be achieved.

Today the social media sites present the most up-to-date information and buzz about current events. This data contains a great amount of valuable information that can be helpful for marketing insights from the perspective of the organizations. Many of the tweets posted daily are either of limited interest or redundant which results in information overload. Here lies the opportunity to benefit from more structured representations of events that are synthesized from individual tweets.

Event Detection is the process of detecting stories about events in a stream of social media data produced by social media websites. A possible application for this task is to alert a news analyst when a new event occurs, e.g., an aeroplane crash, an earthquake, governmental elections, etc. Twitter serves as one of the distribution outlet for those looking to share their personal news and interests. So it hosts substantial amounts of user-contributed textual content. However because this data is in massive scale and heterogeneous, identifying events on Twitter is challenging.

1.1 PESTLE Framework

By leveraging the wealth of available social media data, events of different types and scale can be identified and characterized. The goal of our system is to provide the basics that are required to conduct the PESTLE (P: Political, E: Economical, S: Social, T: Technological, L: Legal and E: Environmental) analysis onto the environment. It combines all the representative factors in terms of detected events in one place which are needed to be analyzed based on the current market situation. The PESTLE framework is designed to provide organizations with an analytical tool to identify various macro-environmental factors that may affect business strategies as shown in Fig. 1. It also helps to assess how these factors may influence business

performance now and in the future (http://pestleanalysis.com). The PESTLE Framework includes six types of environmental influences: political, economical, social, technological, environmental and legal.

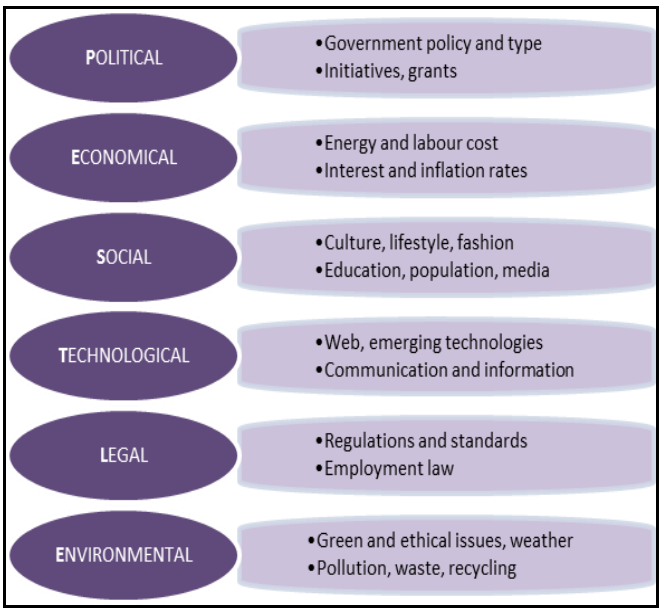

Fig -1: PESTLE Framework

In conducting a PESTLE analysis, organizations may create strategies that take several macro-environmental factors into consideration, so that the strategy formulation process will be as sensitive to current and future environmental factors as possible. The results produced by our system can be the main drivers of change that may affect business strategies, and the factors that are most likely to influence the performance of the business.

2. LITERATURE REVIEW

Earlier work on event detection in textual news (e.g., newswire, radio broadcast) [1], [2], leveraged natural language processing tools (e.g., named-entity extraction, part-of-speech tagging) for on-line identification of news events in a stream of data. In [3], the author surveys the approaches and algorithms that can be used to detect emerging trends from both text streams and archives which result in efficient, organized and classified information for the user of today's era. New Event Detection (NED) is one of the five tasks in the Topic Detection and Tracking evaluations performed each year by the National Institute of Standards and Technology. NED is especially useful for the task of scanning multiple news sources for the latest news. It can be used in a categorization system to identify new categories of news stories and these stories can be used as examples of new categories. People who need to know the latest news when it happens, such as government analysts or financial analysts and stock market traders, can use New Event Detection to more quickly identify new events. To address this, Zhang et al. [4] proposed use of news indexing-tree dynamically whereas Brants et al. [5] used natural language processing. The work of [6] extracted meaningful semantic features such as names, time references, and locations, and learned a similarity function that combines these metrics into a single clustering solution. A lot of research papers explore the unknown event identification scenario in social media. Weng and Lee [7] proposed wavelet based signal detection techniques for identifying real-life events on Twitter which can detect significant bursts or trends in a Twitter data stream. Breaking news events on Twitter are identified using clustering, along with a text-based classifier and a set of news seeders by [8] which are handpicked users known for publishing news. [9] detects events by exploring their textual and temporal components using hierarchical clustering. Sakaki et al. [10] developed technique for identifying earthquake events on Twitter by monitoring keyword triggers (e.g., earthquake or shaking). Clustering techniques are also employed by [11] for Twitter based event detection and analysis system. They focus on Crime and Disaster related Events (CDE), such as shooting, car accidents, or tornado. Authors in [12] make use of Support Vector Machine(SVM) and incremental clustering for social event detection. A method to automatically detect and identify events from social media sharing web sites is proposed by [1]. They use EventMedia dataset which consists of images and videos. The work of [13] has contributed by performing open domain extraction of events using natural language processing. The researchers in [14] propose locality sensitive hashing, classification, boosting, information extraction and clustering for extracting local news events from Twitter. The work of [15] implement and use a combined log-likelihood ratio approach for the geographic and time dimension of real-life Twitter data.

A segment-based event detection system for tweets, called Twevent is discussed by Li et al. [16]. It initially detects tweet segments that are bursty as event segments and then clusters the event segments into events considering both their frequency distribution and content similarity. Detection of social events using robust high-order co-clustering is done in [17]. The authors in [18] make use of content extraction and aging theory as part of topic detection technique. A user interactive system named TwitterMonitor, which performs trend detection on Twitter messages by first detecting the bursty keywords and then applying context extraction algorithms for text analysis is presented in [19]. The work of [20] contributed by identifying real and non-real world event identification using post clustering classification. In [21] researchers have developed event detection approach that allows to retrieve the most emergent topics for organizations. Their aim is to identify them before they become hot topics using Support Vector Machine. This aim is different from ours because we propose the classification of detected events based on PESTLE model. For this purpose we make the use of Naive Bayes classifier.

3. SYSTEM ARCHITECTURE

For the implementation of the system we elected to use an incremental clustering algorithm to effectively cluster a stream of Twitter messages. It detects a suitable cluster assignment based on the messages similarity to existing clusters. This forms the clusters of the tweets which are

similar with respect to a particular topic. The architecture of the system is as shown in Fig. 2. The input dataset consists of Twitter messages. For the training and testing phases of event identification we use human annotators to label clusters. These labelled clusters are used as training set and test set. We then make use of Naive Bayes classifier that is trained to distinguish among various types of clusters to detect the events and classify those.

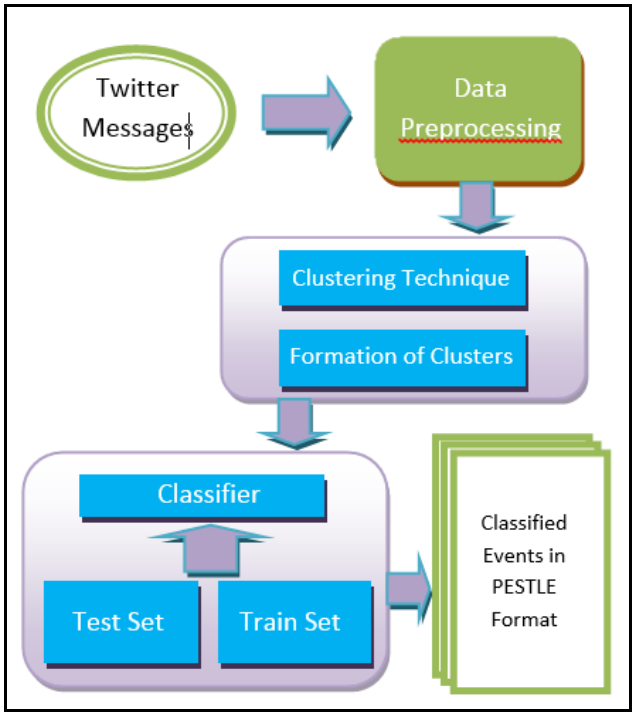

Fig -2: System Architecture

3.1 Twitter Dataset

Twitter has given a new dimension to the way messages are formulated, published, and distributed around the web. It offers freedom for users to generate messages in a quick and easy way without any rules, regulations. The only limitation comes in the form of the message length which is of 140 characters. Twitter users post messages with a variety of content types, including personal up-dates and various bits of information. But because of this, these messages contain little textual information, and often exhibit low quality. As a result Twitter data requires lot of pre-processing. The dataset used for our work is provided at http://help.sentiment140.com

3.2 Data Preprocessing

Data pre-processing is an important step in the data mining process. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: 100), impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or data is noisy and unreliable, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. For the event detection system, the dataset used is that of Twitter which is extremely noisy. Data preprocessing mainly consists of cleaning or filtering of this noisy data. As a part of that process, removal of URLs (e.g. http://example.com), Twitter user names (e.g. @alex with symbol @ indicating a user name), Twitter special words (such as RT6) and emoticons has been done. The stopwords and articles (a, an, the) are also removed from the dataset.

3.3 Clustering

Twitter data evolves constantly. In such scenario it is important to consider a clustering algorithm which is capable of handling continuous stream of data. We make use of an incremental clustering algorithm [22] having a threshold τ. Such a clustering algorithm considers each tweet in turn and determines the suitable cluster assignment based on a similarity function. It computes its similarity (di; Cj) against each existing cluster cj. If no cluster has the similarity more than τ, the algorithm generates a new cluster for that tweet. Otherwise it is assigned to an existing cluster with maximum similarity. The centroid of a cluster is the average weight of each term across all documents in the cluster. Each tweet is represented as a tf-idf weight vector depending on its textual content. As the clustering similarity function, the cosine similarity metric is used which is defined in [23]

$$sim(d,d') = \frac{\sum w\ weight(w,d) * weight(w,d')}{\sqrt{\sum w\ weight(w,d)^2}\sqrt{\sum w\ weight(w,d')^2}} \quad (1)$$

Where

$$weight(w,d) = tf * idf$$

The hashtag terms often indicate the contents of the message, so their weight is doubled.

3.4 Event Classification

We compute the features of Twitter message clusters in order to reveal characteristics that will help detect clusters which are associated with events. While each of these features individually may not indicate event content, combining them with other revealing features (e.g., using a trained classifier) can help identify event clusters. The basic approach is to employ text classification which can identify events based on the messages contained in the cluster. This is achieved by training the Naive Bayes classifier which considers all messages in a single cluster as a single document. As features it incorporates the tf-idf weights of message contents. This classifier that distinguishes between various kinds of events is described in [8] to detect news in Twitter messages. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem. It improves the tasks of the web mining by accurately

classifying documents [24]. It has the advantage that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. The probability model for a classifier [25] is a conditional model

$$p(C|F1, F2, \ldots . F_n) \quad (2)$$

over a dependent class variable C with a small number of outcomes or classes, conditional on several features F1 through Fn.

Using Bayes' theorem [25],

$$p(C|F1, F2, \ldots . F_n) = \frac{p(C)p(F1, F2, \ldots . F_n|C)}{p(F1, F2, \ldots . F_n)} \quad (3)$$

The denominator is independent of C and the values of the features Fi are given such that the denominator is effectively constant. So the interest lies only in the numerator of the fraction in practice.

We train the classifier to distinguish between the type of event clusters i.e. political, economical and so on. Prediction is done by this classifier about which clusters correspond to events. For the identification of event clusters, features of all clusters are computed. Then the classification model is used with each cluster's feature representation to predict the probability that the cluster contains event information. Event and event related terms are specified using tf-idf as follows:

$$tf - idf = tf(pw) * idf(pw) \quad (4)$$

$$tf - idf = tf(pw) * idf(pw)$$

where, pw are political words and tf(pw) is the word count of pw from total word count; idf(pw) is the total word count out of political word count pw.

Table -1: Sample Events in PESTLE format

Event Type	Detected Events	Sample Terms
Political	US Government Shut Down	Government, Shut Down, Justice, Marriage, Act
Economical	Sanctions against Iran	Bank, Sanctions, Oil, Industry
Social	London Olympic	Olympic, London, Games, Oscar
Technological	Apple iPhone5 Launch	Launch, Apple, Big data, iPhone, Business Intelligence
Legal	30th International Financial Law Conference	Dispute, Resolution, Law, Conference, Legal
Environmental	Drought in North America	Drought, Flood, Global warming, Volcano

Table -2: Performance Measurement

Precision	Recall	F1
0.89	0.80	0.84

4. RESULTS

The output produced at this stage of the system is in the form of events and their associated terms as shown in Table 1. The performance measure in terms of precision, recall and F1 measure is reported in Table 2 which we compare with the performance of TWICAL [13] in Fig. 3.

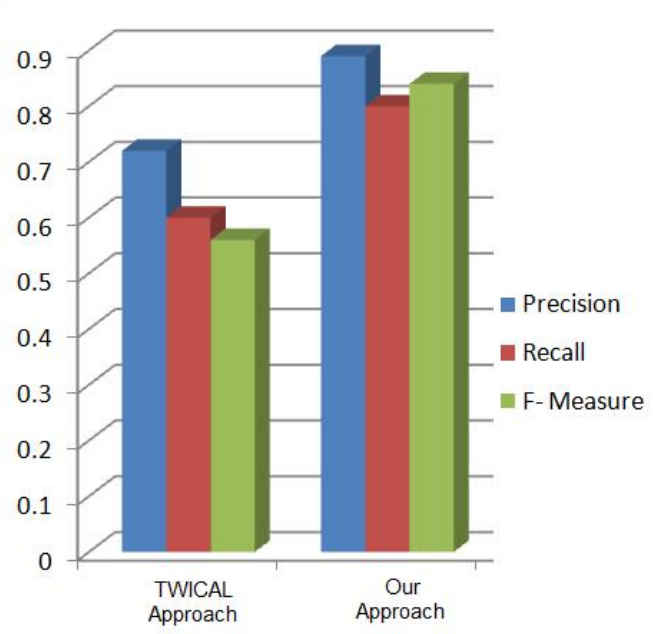

Fig -3: Performance Comparison with TWICAL

5. CONCLUSION AND FUTURE SCOPE

Social media data contains valuable information hidden inside which offers ample opportunities for social media mining to discover actionable knowledge. We proposed a system that detects events from social media site i.e. Twitter. The Twitter data comes in unstructured and noisy format. The use of incremental clustering is made for the formation of clusters of data items which are closely related. This algorithm does not need a priori knowledge of number of clusters and is also scalable. We used Naive Bayes classifier to predict which clusters belong to an event. Finally events were identified and classified in PESTLE frame-work. We further wish to extend our work by classifying the messages into events such as ethical, demographical etc. which will give a detailed overview of more macro-environmental factors. A richer representation of events can also be achieved by classifying entities in relation to their roles in a specific event.

ACKNOWLEDGEMENTS

We would like to thank the publishers, researchers for making their resources available. We also thank the college authority for providing required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1] X. Liu, R. Troncy, and B. Huet, "Using social media to identify events," in Proceedings of the 3rd ACM SIGMM international work-shop on Social media. ACM, 2011, pp. 3–8.

[2] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," 1998.

[3] M. W. Berry and M. Castellanos, Survey of text mining, 2004.

[4] K. Zhang, J. Zi, and L. G. Wu, "New event detection based on indexing-tree and named entity," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007, pp. 215–222.

[5] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003, pp. 330–337.

[6] J. Makkonen, H. Ahonen Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," Information Retrieval, vol. 7, no. 3-4, pp. 347–368, 2004.

[7] J. Weng and B.-S. Lee, "Event detection in twitter." in ICWSM, 2011.

[8] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2009, pp. 42–51.

[9] R. Parikh and K. Karlapalem, "Et: events from tweets," in Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013, pp. 613–620.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in Proceedings of the 19th international conference on World wide web. ACM, 2010, 851–860.

[11] R. Li, K. H. Lei, R. Khadiwala, and K.C. Chang, "Tedas: a twitter-based event detection and analysis system," in Data Engineering (ICDE), 2012 IEEE 28th International Conference on. IEEE, 2012, 1273–1276.

[12] Y. Wang, H. Sundaram, and L. Xie, "Social event detection with interaction graph modeling," in Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 865–868

[13] A. Ritter, O. Etzioni, S. Clark et al., "Open domain event extraction from twitter," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 1104–1112.

[14] P. Agarwal, R. Vaithiyanathan, S. Sharma, and G. Shroff, "Catching the long-tail: Extracting local news events from twitter." in ICWSM, 2012.

[15] A. Weiler, M. H. Scholl, F. Wanner, and C. Rohrdantz, "Event identification for local areas using social media streaming data," in Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks. ACM, 2013, pp. 1–6.

[16] C. Li, A. Sun, and A. Datta, "Twevent: Segment-based event detection from tweets," in Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012, pp. 155–164.

[17] B.K. Bao, W. Min, K. Lu, and C. Xu, "Social event detection with robust high-order co-clustering," in Proceedings of the 3rd ACM conference on International conference on multimedia retrieval. ACM, 2013, pp. 135–142.

[18] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM, 2010, p. 4.

[19] M. Mathioudakis and N. Koudas, "Twittermonitor: Trend detection over the twitter stream," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '10. New York, NY, USA: ACM, 2010, pp. 1155–1158. [Online]. Available: http://doi.acm.org/10.1145/1807167.1807306

[20] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter." in ICWSM, 2011.

[21] Y. Chen, H. Amiri, Z. Li, and T.S. Chua, "Emerging topic detection for organizations from microblogs," in Proceeding of 36th International ACM SIGIR Conference on Research and Development in Information Retrieval

[22] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in Proceedings of the Third ACM International Conference on Web Search and Data Mining, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 291–300. [Online]. Available: http://doi.acm.org/10.1145/ 1718487.1718524

[23] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '04. New York, NY, USA: ACM, 2004, pp. 297–304. [Online]. Available: http://doi.acm.org/10.1145/1008992.1009044

[24] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," 2007.
[25] H. Zhang, "The optimality of naive bayes," A A, vol. 1, no. 2, p. 3, 2004.

BIOGRAPHIES

Vaishali Ugale: Received the BE degree in Computer Engineering She is pursuing Post Graduation in Computer Engineering from Savitribai Phule Pune University of Pune at D. Y. Patil College of Engineering, Akurdi, Pune.

Ms. Soudamini Pawar: Received the ME degree in Computer Engineering from Savitribai Phule Pune University of Pune. She is currently working as an Assistant Professor and PG Coordinator in D. Y. Patil College of Engineering, Akurdi, Pune.