Topical Evolution and Regional Affinity of Tweets

Lipika Dey, Arpit Khurdiya, Diwakar Mahajan
Innovation Labs, Delhi
Tata Consultancy Services
Gurgaon, India
{lipika.dey, arpit.khurdiya, diwakar.mahajan}@tcs.com

***Abstract*— Business organizations are increasingly showing interest in Twitter content to know their consumers. Tracking popular tags and trends give some idea about what people are talking about. However, in order to act on the knowledge acquired, they need more detailed information like regional variability in content, exact location of discontent if any, regional affinities and influences etc. In this work, we present methods to identify topics of discussion in tweets using a LDA-based approach, which can identify emerging or evolving topics. Regional analysis of topics can provide interesting business insights about consumer expectation or behavioural variations. Further, regional distribution of topics are analysed to identify clusters of regions that tend to behave similarly over extended periods of time.**

Keywords- Topic extraction, regional clusters, Social Media Analytics, Clustering, Regional Dispersion

I. INTRODUCTION

The micro-blogging platform Twitter has emerged as one of the most commonly used social platforms, over which users exchange information across continents. While there are a multitude of commercial products providing tools or services to bring relevant Twitter content to the business users' desktops, they mostly present frequencies of words, hash-tags, user-mentions, trends etc. Interpretation, aggregation and consumption of social-media content still remain human-centric tasks, thus introducing subjectivity, instability and variability into the results.

In this paper we propose methodologies to analyze twitter content and provide meaningful insights that can be channelized to a business analytics framework. We present measures and methods that can be used to convert social media content into meaningful complex aggregates like topics or events. Insights are provided as emerging topics, topic evolution and topic-spread across regions. It may be further possible to use topic significance in a predictive framework to predict impact of Twitter content on business.

The rest of the paper is organized as follows. We present a brief overview of earlier similar work done on Twitter analysis in section 2. Section 3 presents the outline of a prototype system built to collect and analyze large volumes of tweets in areas of interest, along with functional descriptions of the components. Section 4 presents the analytical framework. Section 5 presents some results obtained from experiments. Finally, we conclude with future directions to extend the ongoing work in section 6.

II. SURVEY OF EARLIER WORK

TweetMotif [1] presented an unsupervised approach to identify topics from tweets based on message clustering. In [2], it was proposed that hash-tags that appear in tweets can be viewed as approximate indicators of a tweet's topic. Agarwal et al. [3] presented real-time techniques to discover events from microblog message streams by modeling the problem as that of discovering dense clusters in highly dynamic graphs. Another area of interest while analysing social media content has been to understand the factors that affect content diffusion. Yang and Leskovec [4] identified four major factors that affect content diffusion. These are personal preferences of the users, their immediate network of friends on the network, geographic or regional issues and events and world-wide happenings. Due to the enormity of the content in social media, most models for content diffusion consider isolated factors and their interactions [5], [6], [7]. In [8], a combined model was presented to analyse the influences of various factors affecting individual messages posted in social media. Eisenstein et al. [13], focused on predicting the geo-location of a tweet based on the text in the tweet, which made use of the geo-tagged information in the tweets as the gold standard label for measurement.

III. DATA ACQUISITION, PRE-PROCESSING AND TOPIC EXTRACTION

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this template and download the file for A4 paper format called "CPS_A4_format".

Fig. 1 illustrates the different functional components of the proposed analytical framework. Tweets are collected using Twitter4J API that gathers twitter streams based on specified keywords. Data processing and churning is performed over a scalable architecture implemented using NoSQL database namely MongoDB and Hadoop. Indexing and retrieval are aided by SOLR. All acquired tweets are stored in the repository along with their meta-data like author profile, time of tweet and origin of tweet, if available. The location disambiguation component is responsible for assigning a location to the tweet based on the associated latitude-longitude information for geo-tagged tweets. For those that are not geo-tagged, user profile, time-zone etc. are used to assign a possible location. The accuracy for the assignment is around 72%, which is similar to results

978-0-7695-5066-4/13 $26.00 © 2013 IEEE
DOI 10.1109/ISCBI.2013.66

reported by Hecht et.al [12]. The de-duplicator component identifies duplicates and near-duplicates and groups them together into buckets. Each bucket is associated with a representative tweet, which is the one with the smallest time-stamp. The association-analyzer analyzes the representative tweets of each bucket to identify possible number of distinct topics and also identify seed terms for topic extraction. Map-Reduce framework is used to construct term-document matrix. Mapper tokenizes the tweets and generates a token frequency map corresponding to each document. Reducer aggregates them to build the matrix, which is used by the topic-extractor.

Though theoretically, re-tweets are supposed to be duplicates, in reality, these are often near duplicates which contain few additional words. We employ Min-Hash Clustering [10] to collect all near duplicate tweets into a single group. It is observed that the total volume of tweets is reduced to about 40% of the original volume.

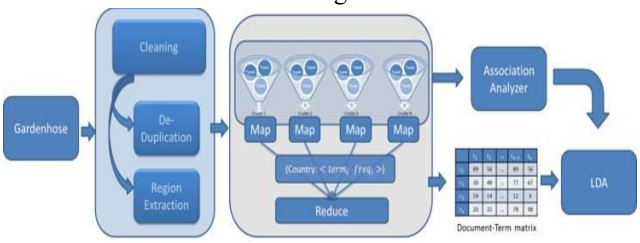

Figure 1. Information Processing Architecture

IV. Topic-based Analytics to derive Business Insights

Twitter data can provide business organizations critical insights regarding changing market landscape, consumer expectations, evolving threats and opportunities etc. The proposed analytical methods are aimed at automating the derivation of key insights.

A. Discovering New Topics

Topic extraction identifies sets of topics that are discussed across the globe over a specified period of time. We have used the standard LDA [9] based generative model for topic extraction. A document collection is modeled as a distribution of topics while topics are modeled as distribution of words. Each document can be modeled as a probability distribution over a pre-defined number of topics. Correlation of topics over two time-periods can help in identification of new topics. For each given day i let T_i denote the set of topics generated by the algorithm and let t_i^k denote one of the topics in T_i. Pearson correlation coefficient C_{xy} between any two topics t_i^x and t_j^y for two given days i and j respectively is computed as

$C_{xy} = Corr(t_i^x, t_j^y)$, where $Corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, x and y are the word-probability distributions associated to t_i^x and t_j^y respectively. A topic t_i^x is identified as emergent or new if its correlation with all topics for a pre-specified number of earlier days is found to be less than a specified threshold. The threshold is kept at 0.6. A topic is assumed to be an evolution of another topic from earlier day if the two are highly correlated.

B. Finding Regional Affinities

Topics can also be associated to regions. Clustering the topic-region space identifies regional affinities for topics. Long term regional affinities are identified using a meta-clustering based approach over the regional clusters obtained for each day. Since all topics are not equally significant and on any given day there are a large number of topics discussed globally, we first applied singular value decomposition (SVD) on the region-topic matrix, to identify only those topics which play significant role in distinguishing among regions. The process is explained below.

- Finding regional affinities for each day
  - For each day d
    - Let M be a $r \times k$ matrix with rows representing regions and columns representing topics.
    - Obtain Singular Value Decomposition of M.
    - Consider $\widetilde{M}$ as an approximation of M containing top p singular-values of M.
    - Apply k-means clustering on entities represented in rows $\widetilde{M}$ to obtain groups of regions that discuss similar topics.
- Meta-clustering
  - Build Region adjacency matrix A as follows
    - Two regions are considered to be adjacent to each other if they belong to the same cluster for a given minimum number of days.
    - Perform Edge-between-ness clustering [11] on A.
    - Output – long-term clusters of regions.

Long term regional affinities are very useful business analytical tools. When a business critical topic is found to emerge at a region, regional affinities can be used to predict its spread in other regions.

V. Experiments and Results

We now present results from a collection of approximately 10 million tweets related to Apple devices like iPhone, iPad etc.

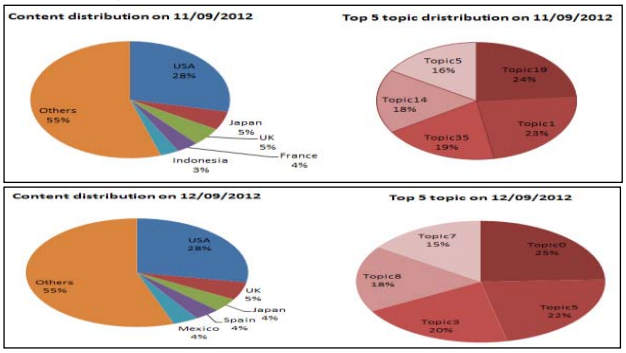

Figure 2. Region-wise distribution shows changing role of countries on 11th and 12th September.

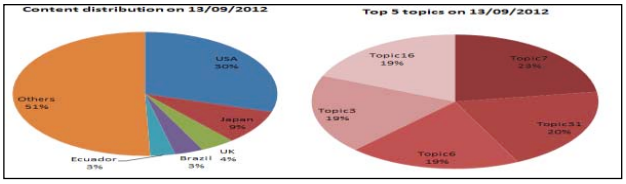

Figure 3. Region-wise distribution shows changing role of countries on 13th September.

Figure 2 and 3 shows how different countries participated on iPhone related discussion on 11th, 12th and 13th September. Figures 4 shows content of popular topics. Majority of the discussions on 11th Sept. and 12th Sept. were around intention to own an iPhone or the launching ceremony. On 13th Sept. people started tweeting about its looks.

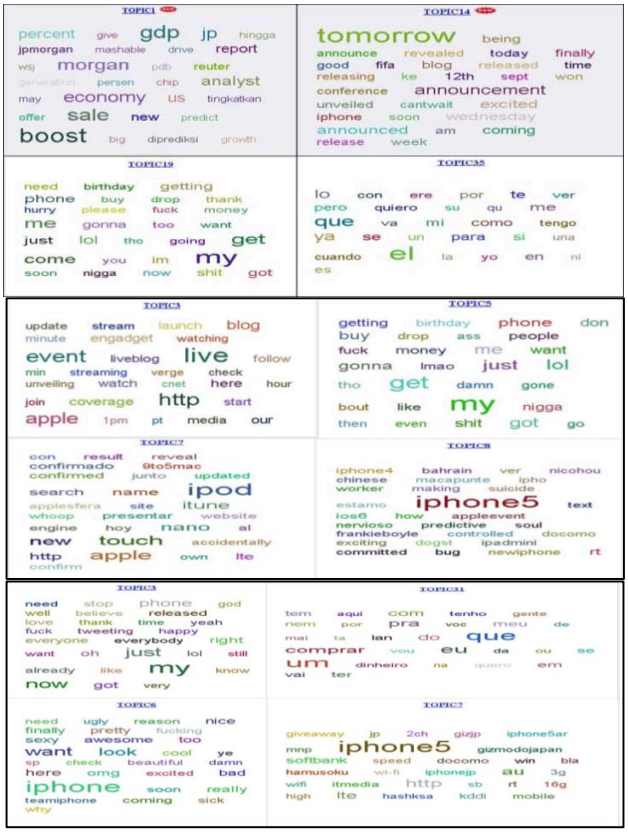

Figure 4. (a) On 11/09/2012, tweets were about buying it next day. (b) On 12/09/2012, it was about the launching ceremony. (c) On 13/09/2012, tweets were about looks of iPhone.

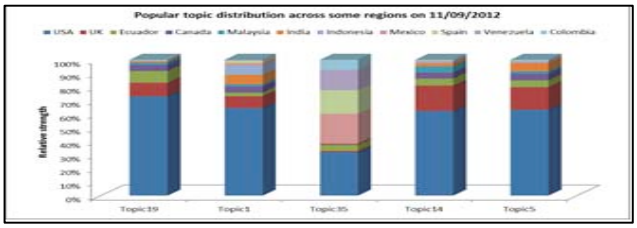

Figure 5. Regional distribution of popular content related to iPhone on 11September 2012

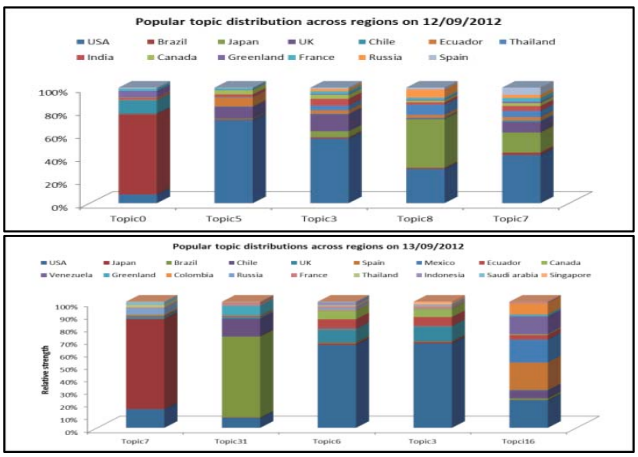

Figure 6. Regional distribution of popular content related to iPhone on 12th & 13th September 2012

Figure 5 and 6 shows that while USA was always the largest participant, participation of Brazil, Canada, Japan, and Spain graudally increase. It was found that while other regions mostly talk about buying an iPhone or its cost, tweets in Japan were around technology aspects like high-speed, wifi or softbank etc. Figure 7 presents results for topics that emerge around three key competitors of iPhone – Samsung, Nokia and Blackberry, which had around 50000, 22000 and 9000 tweets respectively. In case of Samsung, top content was around the patent battle between Apple & Samsung and a public dig at iPhone in an advertisement by Samsung respectively. Feature-wise analysis for Nokia reveals that "better" was used for Nokia Lumia 900 and "underwhelms" was used for iPhone5. However, since the number of tweets is only 400, it is obvious that Nokia Lumia is no competitor for iPhone.

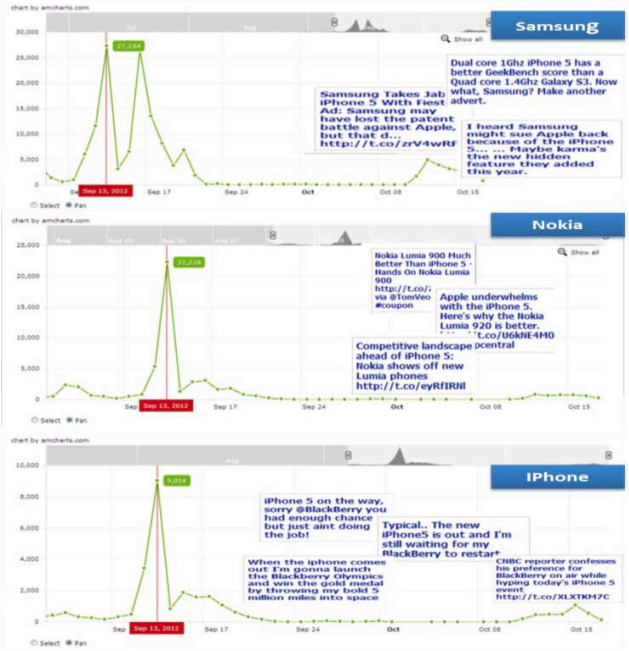

Figure 7. Timeline and top tweets for Samsung,Nokia & Blackberry

Figure 8 shows Region-wise distribution for topics discussing the key competitors namely Samsung, Blackberry & Nokia. While Samsung has world-wide popularity, it's presence is strong in South East Asia. Blackberry has least regional spread. Nokia has less presence in South East Asia.

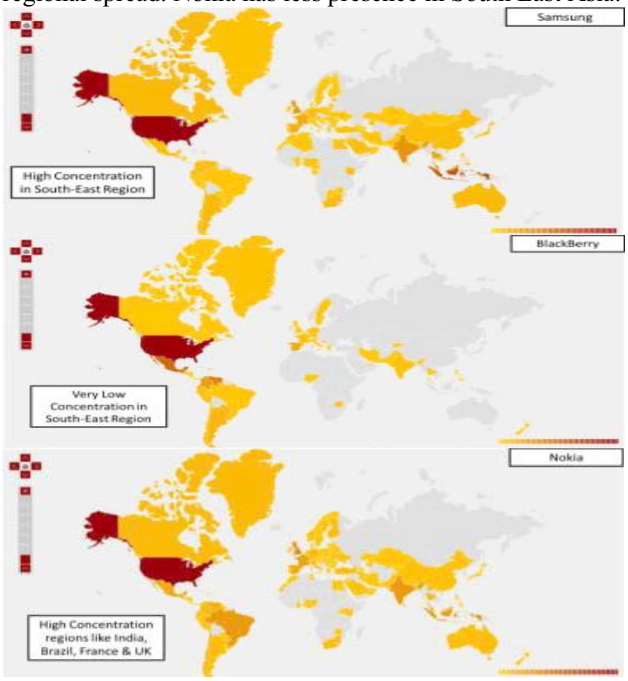

Figure 8. Regional distribution of tweets on Samsung, Blackberry & Nokia respectively

Figure 9 presents some long-term clusters of regions that tweet on same topics. The clusters distinctly show that geographical distance and language play major role in bringing geographically distant countries together. Thus Brazil and Portugal tweet similar content in Portugese though they are geographically apart. Similarly, Mexico and Spain are always in the same cluster. Strong clusters are observed for middle-east Asian and European countries also.

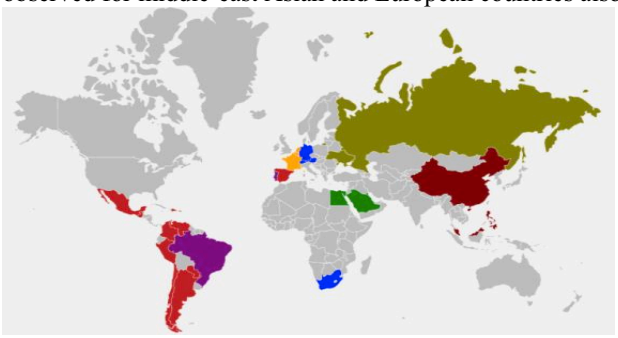

Figure 9. Six Long-term clusters of regions obtained for the iPhone domain

VI. CONCLUSION

In this paper, we have presented methods to analyze Twitter content to obtain business critical knowledge. We have shown that topic-based analysis can provide more semantic cue about consumer discussions. We also presented methods to identify regions that tweet similar topics. It was further shown that there is a natural tendency for geographically close countries or countries speaking same language to tweet on similar topics. Regional affinity for topics can be effectively utilized by organizations to take pro-active steps to prevent spread of damaging topics provided they are detected early enough. Presently this work is extended to automatically classify extracted topics into different categories like consumer expectations, product issues, service issues, comparison with competitor product etc. Further, we intend to model each of these topic classes explicitly to understand how different types of content spread.

REFERENCES

[1] B. O'Connor, M. Krieger and D. Ahn, 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*.

[2] Rosa, Kevin Dela, Shah Rushin, Lin Bi, Gershman A. and Frederking R., Topical Clustering of Tweets, SWSM'10, July 28, 2011, Beijing, China.

[3] Agarwal M. K., Ramamritham K. and Bhide M., Real Time Discovery of Dense Clusters in Highly Dynamic Graphs: Identifying Real World Events in Highly Dynamic Environments, Proceedings of the VLDB Endowment, Volume 5, Issue 10, June 2012, pp. 980-991.

[4] J. Yang and J. Leskovec. Patterns of temporal variation in online social media. In WSDM, 2011.

[5] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In International AAAI Conference on Weblogs and Social Media, 2010.

[6] Z. Wen and C. Lin. On the quality of inferring interests from social neighbors. In SIGKDD, 2012, pages 373–382.

[7] A. Ahmed, Y. Low, M. Aly, and V. Josifovski. Scalable distributed inference of dynamic user interests for behavioral targeting. In SIGKDD, 2011.

[8] Lakkaraju H., Bhattacharya, I. and Bhattacharyya., C., Dynamic Multi-Relational Chinese Restaurant Process for Analyzing Influences on Users in Social Media, Proc. ICDM 2012

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

[10] Broder, A. Z. "On the resemblance and containment of documents, in 'Compression and Complexity of Sequences'" , (1997), IEEE Computer Society Press, Salerno, Italy , pp. 21—29

[11] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (2002) 7821–7826.

[12] Brent Hecht , Lichan Hong , Bongwon Suh , Ed H. Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. Proceedings of the Annual Conference on Human Factors in Computing Systems, May 07-12, 2011, Vancouver, BC, Canada.

[13] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1277–1287, Cambridge, MA, October. Association for Computational Linguistics002E