Twitter Based Hyperlocal News Extraction

[1] Prof.Sunayana V Jadhav, [2] Rahul Ahuja, [3] , Mukesh Wadhwa, [4] Dharmish Bhatt, [4] Alekh Shah
[1]Assistant Professor, [2],[3], [4], [5], Final Year Engineering Students
[1]Infirmation Technology,
[1] K.J.Somaiya College of Engineering,
Vidyavihar, Mumbai 4000072,India.

Abstract: Life is amorphous jumble of events, plummeting over each other, shoving into other. Everyday Journalists make sense of this chaos so that the general public receives it as a story, a neatly packages story. But while deciding which story is worth publishing a lot of them never make onto the news outlet because their coverage is small or it affects a small amount of people , sometimes a small community. With the dawn of online news, we have connected the world but somehow disconnected the locality. We intend to bring community level hyperlocal news stories to our audience depending upon spatial attributes of our news and our users..

Index Terms **- Hyper local, Recommendation system, Feature extraction.**

I. INTRODUCTION

Proposed system intends to deliver hyper local news to community users in a swift manner ,considering its location relevance. We intend to make use of News channels RSS feeds and other non-professional news outlets such as twitter to capture our data . Twitter has proved to be an awesome micro- blogging platform where even the official authorities such as the State and Central Government have made critical announcements This data is to be then analyzed to extract relevant information. This information includes the location , type of news and other factors which helps us to measure the relevance of the news to a particular user. When we talk about relevance of news, the more it tops locality of an individual, its relevance to masses decreases [1].

II. EXISTING WORK

There are already some major implementations of hyperlocal news publishing system. Blockfeed App, a New York startup established in year 2015. Block feed provides a news feed tailored to user's exact location in the city . Blockfeed goes around hundreds of stories from local news blogs, news outlets and caters news to users depending upon their location. Another American startup, Ripple also provides hyper local news from crawling local news platforms and blogs. It allows registered news hunters to publish news on the Ripple too. The news appear alongside professionally edited news.[1]

2.1 Local News Chatter: Augmenting Community News by Aggregating Hyperlocal Microblog Content in a Tag Cloud

Kyungsik Han, Patrick C. Shih, and John M. Carroll
College of Information Sciences and Technology, Pennsylvania State University, University Park, Pennsylvania, USA

Being aware of local community information is critical to maintaining civic engagement and participation. The use of online news and microblog content to create and disseminate community information has long been studied. However, interactions in the online spaces dedicated to local communities tend to only garner very limited usage, and people often do not consider microblog content as a meaningful source of local community information. Local News Chatter (LNC) was designed to address these challenges by augmenting local news feeds with microblog content and presenting them in a tag cloud that displays news topics of varying popularity with different tag sizes. Our study with 30 local residents highlights that LNC increases the visibility of hyperlocal community news information and successfully utilizes microblog as an additional information layer.

LNC also increases one's community awareness and shows the potential for leveraging community knowledge as a deliberation platform for local topics[2]

2.2 Automatic Classification of Disaster-Related Tweets

Beverly Estephany Parilla-Ferrer, Proceso L. Fernandez Jr., PhD, and Jaime T. Ballena IV, PhD

The social networking site Twitter has become one of the quickest sources of news and other information. Twitter information feeds known as tweets, are voluntarily sent by registered users and reach even non-registered users, sometimes ahead of traditional sources of mass news. In this study, we develop some machine learning models that can automatically detect informative disaster-related tweets.

A dataset of tweets, collected during the Habagat flooding of Metro Manila in 2012, was used in building the classifier models. A random subset of this dataset was manually labeled as either informative or uninformative to produce the ground truth. Two machine learning algorithms, Naive Bayes and Support Vector Machine (SVM), were used to build models for the automatic classification of the tweets, and these models were evaluated across the metrics of accuracy, precision, recall, area under curve and F-measure. Experimental results show that the model generated from SVM has significantly better results compared to that of the Naive Bayes.

This study also revealed that uninformative tweets outnumbered informative tweets, suggesting that the subscribers used Twitter to broadcast more of tweets that express their subjective messages and emotions regarding the Habagat event.

However, the informative tweets were more likely to be retweeted than uninformative tweets, indicating that subscribers retweet messages they deem informative and useful for public awareness.

These insights, together with the built classifier models, can help in the development of a system that can sift through the voluminous Twitter data and in real-time detect informative disaster-related tweets so that appropriate action may be done promptly.[3]

III. Proposed system

We intend to make use of micro-service architecture. One micro-service (News Spider) will be responsible for gathering all the news from different outlets such as RSS feeds, micro-blogs, local blogs , etc and push it onto a queue to be processed. The queue is known as News Pool. The information present in our News Pool is processed by NEEX (News Entity EXtractor). NEEX is a micro-service that fetches a raw news from News Pool and makes use of text processing techniques such as lemmatization/stemming, Named Entity Recognition, Sentiment Analysis. It is the job of NEEX to identify relevant information and filter out garbage tweets and information from our relevant required news information NEEX creates an Attribute Rich News Object (ARNO). ARNO is a complete news content which is ready to be consumed by News Curator Component (NCC).ARNO consists of the location of news, type of news, it's source and other temporal attributes. From user profile perspective, we intend to monitor our user's locations which would enable us to create a User Profile. User Profile is used by NCC to push ARNO to our target audience.

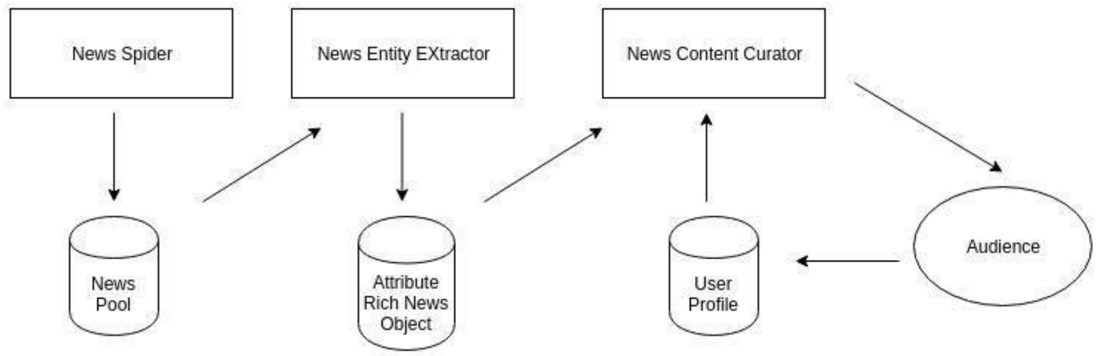

Figure 1.1: Block diagram for proposed system.

News Pool shall make use of NoSQL database as it is best suited for unstructured data. On the other side, an RDBMS would be a better choice for ARNO.

3.1 Dataset for News Pool

A dataset containing raw textual information , RSS feeds, tweets and all other noise that exists on the internet which will be extracted by our News Spider.

3.2 Dataset ARNO

A collection of Attribute Rich News Object which would be the output of NEEX. This information will be used by our NCC to curate content in conjunction with User Profile.

3.3 User profile Dataset

A user profile dataset containing the users general information , spatial attributes and content preference can be made. This dataset will act as an input to our NCC which in conjunction with our ARNO will be used to curate content for users.

IV. IMPLEMENTATION

Every news channel has an RSS feed outlet which acts as a radio and transmits news in XML format. Our system intends to subscribe to multiple such RSS feeds and extract happenings from them. We also intend to crawl social media platform such as twitter. On the news data collected from these platforms and channels, entity extraction and sentiment analysis can be performed. i.e. The unstructured information in its natural textual form is parsed and broken down into defining elements. This newly found object is then stored into our news store , which is a database. Our android application tracks user's movement and creates a user profile depending upon his frequently visited locations. This enables us to push relevant news objects to our users. The news parsing and analysis makes use of NLP to extract entity (location and object of news, type , etc)

V. ADVANTAGES

- Personalized News Content
- Relevant location specific news
- Revenue generation through local advertisements
- Scope for local journalists
- Platform for students pursuing journalism

VI. TOOLS AND TECHNIQUES

- MongoDB – For storage of unprocessed data.
- SQL Database – For storage of processed data.
- GitHub – For virtual sharing of source code.
- Web Crawler – who will fetch the news and give to the application for processing.
- RSS Feeds and Twitter API – Used to fetch the news.
- Java 8 – For development.
- jUnit and Jmeter – Testing.
- Jdk 1.8
- android-studio-ide-181.5014246

VII. CONCLUSION

The proposed hyperlocal news curation system will povide a personalized content curated upon personal preference, spatial attribute , temporal feature and relevance to an individual. This can form the basis of a revenue model where we can give a new outlet to local businessman for advertisement. It will also promote hyperlocal journalism and engage the common masses in news reporting, hence improving civic involvement.

VIII. ACKNOWLEDGMENT

“Author would like to acknowledge the management of K J Somaiya College of Engineering, for supporting this Research Work”

IX. REFERENCES

[1] Vieweg, S. (2010). Microblogged Contributions to the Emergency arena: Discovery, In- terpretation and Implications. Paper presented at the Computer Supported Collaborative Work. Palen, L., & Vieweg, S. (2008). The Emergence of Online Widescale Interaction in Unexpected Events: Assistance, Alliance & Retreat. Paper presented at the CSCW 2008.

[2] Castillo, C., Mendoza, M., & Poblete, B. (2011, 28 March - 1 April). Information Credibil- ity on Twitter. Paper presented at the WWW 2011, Hyderabad, India.

[3] Yates, D., & Paquette, S. (2011). Emergency Knowledge Management and Social Media Technologies: A case study of the 2010 Haitian earthquake. International Journal of In- formation Management,

[4] Palen, L., K. Anderson, G. Mark, J. Martin, D. Sicker, & D. , Grunwald. A Vision for Technology- Mediated Public Participation and Assistance in Mass Emergencies and Dis- asters, University of Colorado manuscript. 2010.

[5]Yin, J., Lampert, A., Cameron, M., Robinson, B., Power R. (2012). Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems. http://dx.doi.org/10.1109/MIS.2012.6

[5] Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010). Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. Paper pre- sented at the CSCW 2010, Savannah, Georgia, USA.

[6] KeyXtract Twitter Model - An Essential Keywords Extraction Model for Twitter Designed using NLP Tools Tharindu Weerasooriya1# , Nandula Perera2 , S.R. Liyanage3 1# De- partment of Statistics and Computer Science, University of Kelaniya, Sri Lanka 2 Depart- ment of English, University of Kelaniya, Sri Lanka 3 Department of Software Engineering, University of Kelaniya, Sri Lanka .

[8] Automatic Classification of Disaster-Related Tweets Beverly Estephany Parilla-Ferrer, Proceso L. Fernandez Jr., PhD, and Jaime T. Ballena IV, PhD