Asian Journal of Convergence in Technology **Volume V Issue I**
ISSN NO: 2350-1146 I.F-5.11

Social Media Text Mining for Decision Support in Natural Disaster Management in Sri Lanka

K. HJayaniImalka, S. C.Premeratne

***Abstract*— With the popularity of the internet and smart devices, social media is viral today among individuals in almost all the ages which help them to create and share their personal feelings, experiences, ideas, as well as information with others connected to them over a computer, mediated technology. Due to this nature when there are emergencies and natural disasters these social media applications tend to be flooded with content generated from the public who affected, who are looking for their family members and friends, who are looking for information as well as with the people engage in humanitarian activities.Therefore, social media has become the first to generate related information when there is a catastrophic event before any of news sites or government bodies engage in disaster management. This social media content is quick accurate and subjective during disaster situations, therefore,can be used as an asset to reduce risk and build awareness among the public about the disaster as well as to provide decision making support to relief efforts. This research focuses on building decision making support using social media content generated during disaster situations in the Sri Lankan context. Mainly the content will be tweets posted by the public during a natural disaster and consisting of text written in English. Therefore,situational awareness building will be done using text mining techniques in this study since the content is unstructured.**

Keywords—social media,natural disaster management,decision making,text mining,natural language processing, situational awareness

I. INTRODUCTION

Social media is very popular today among individuals in almost all the ages which help them to create and share their personal feelings, experiences, ideas as well as information with others connected to them over computer-mediated technology. Social media services are (currently) Web 2.0 Internet-based applications,and user-generated content is the lifeblood of social media. Individuals and groups create user-specific profiles for a site or app designed and maintained by a social media service and social media services facilitate the development of social networks online by connecting a profile with those of other individuals and groups.[1] As at August 2017, most popular social networking site are Facebook, YouTube, Instagram, Twitter, Reddit, Vine, Ask.fm, Pinterest, Tumblr, Flickr, Google+, LinkedIn,etc. [2].

According to the statistics of TRC[1], Sri Lanka's increased internet connectivity have given a boost to the Sri Lankan's presence in the social media, especially on Facebook, the favorite local online hangout. Starting from mid-2016, the Sri Lankans number on Facebook has increased from a 4 million to 5 million.

1http://www.trc.gov.lk/2014-05-13-03-56-46/statistics.html

Due to this nature of social media services during times of disasters, online users generate a significant amount of data, some of which are extremely valuable for relief efforts. [3] The increasing use of social media, such as Twitter and Facebook, by humanitarian organizations, public authorities,and citizens preparing for and responding to disasters generates vast quantities of information.[4] Crowdsourcing of data during such a disaster can aid in the task of decision making.[5]. The public is yet more active online during disasters, increasingly turning to social media for the most up to date information. Social media, however, are used for more than information seeking or sharing during disasters; the public increasingly expectsemergency managers to monitor and respond to their social media posts. [6] Social media improves situational awareness, facilitates dissemination of emergency information, enables early warning systems, and helps coordinate relief efforts [7].

II. DISASTER AND DISASTER COMMUNICATION

A. *Disaster*

The disaster has been defined in many ways; World Health Organization has defined disaster as any sudden occurrence of the events that cause damage, ecological disruption, loss of human life, deterioration of health and health services, on a scale sufficient to warrant an extraordinary response from outside the affected community or area.

Therefore, disaster management is very important to survive in the case of a natural or a major human-made disaster and can be defined as the organization and management of resources and responsibilities for dealing with all humanitarian aspects of emergencies, preparedness, response and recovery to lessen the impact of a sudden disaster.

B. *Social Media*

Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing,and collaboration, as well as websites and applications dedicated to forums, microblogging, social networking, social bookmarking, social curation, and wikis, are among the different types of social media. [10]

In recent years, Twitter has been used to spread the news about casualties and damages, donation efforts and alerts, including multimedia information such as videos and photos. [3] In responding to disasters, including the 2010 Haiti earthquake, the 2012 Sandy superstorm and the 2013 Boston Marathon bombings, social media was used for relaying information, one and two-way communication, offering/requesting assistance and organizing disaster response. [4]

C. *Text Mining during disasters*

With the massive use of social media during disaster situations incoming information filtering is very much important for situational awareness.55 popular text mining tools with their features are mentioned by Arvinder Kaur and Deepti Chopra [11].

III. Literature Review

Researchers conducted in this area of study can be categorized as

a) Usage of social media during disaster situations
b) Assessment of disasters using information posted on social media
c) Collecting information from social media
d) Processing social media messages in mass emergencies and
e) Building decision support systems using social media data posted during disaster situations.

A. *Use of social media during disaster situations*

[15] Describes possibilities of using social media in natural disaster management. In the paper, they have presented an analysis of communication types in between participants in natural disaster events as well as guidelines for organizing information exchange by social media. They have identified that social media can be used in three ways according to many types of research and those are

1. Preparing for a natural disaster
2. Responding during and immediately after the natural disaster
3. Recovering from the natural disaster

B. *Assessment of disasters, Processing social media messages and assist decision making in mass emergencies*

Twitter is a platform which allows its users to post 140-character messages and to follow messages from any other registered users, therefore that openness place Twitter somewhere in between a purely social network and purely informational network [7]. An assessment done with relates to one of costliest disasters in US history, Hurricane Sandy 2012 to show the relationship between proximity to Sandy's path and hurricane-related social media activities. Further, they have demonstrated that per-capita Twitter activities strongly correlate with the per-capita economic damage inflicted by the hurricane [7].

C. *Opportunities and barriers of using social media for disaster preparedness*

Due to network connectivity during disaster situations, public access to social media data gets limited which is a serious obstacle to research in this space [13]. Data preprocessing is done by most researchers using available techniques according to the type of data they are having and goal of analysis. Natural Language Processing is used since all the data consisting of unstructured textual form in social media [22]. By using NLP toolkit tokenizing, part-of-speech tagging (POS), semantic role labeling, dependency parsing, named entity recognition, and entity linking can be performed according to the publication by M. Imran, C. Castillo, F. Diaz, and S. Vieweg [13].

Twitter messages are brief, informal, noisy, unstructured and often contain misspellings and grammatical mistakes [23]. Due to 140-character limitation twitter users intentionally shorten words by using abbreviations, acronyms, slangs, and sometimes words without spaces [23]. Therefore,special attention must be paidto improve the accuracy of Natural Language Processing due to this informal nature.

Opportunities of using social media during disaster situations include Disaster Risk Reduction and preparedness such as by examining and managing the causal factors of disasters, including reducing exposure to hazards, reducing the vulnerability of people and property and increased preparedness for disaster eventsusing eyewitness accounts and images or videos of the impacted area on social media posted by citizens [4].

Online social networks allow the establishment of global relationships that are domain related or can be based on some need shared by the participants and emergency service agencies are utilizing the power of social media and SMS to instantly broadcast and amplify emergency warnings to the public [15]. The further paper emphasizes the following critical tasks that can be implemented by social media.

a) Prepare citizens in areas likely to be affected by a disaster;
b) Broadcast real-time information both for affected areas and interested people;
c) Receive real-time data from affected areas;
d) Mobilize and coordinating immediate relief efforts; and Optimize recovery activities

D. *Practical Extraction of relevant information*

Multiclass categorization of Twitter messages has been done by using well-known algorithms Naive Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF) in a paper published by Imran and all [23]. In the paper, they have mentioned that training all classifiers has been done by preprocessed data. Following are the preprocessing steps they have followed.

1. First, removing stop-words, URLs, and user mentions from the Twitter messages
2. then Stemming using the Lovins stemmer
3. Using Unigrams and bi-grams as features
4. Using information gain, a well-known feature selection method to select top 1k features.

Labeled data used in this task has been annotated by the paid workers [23].

IV. Approach for Social Media Text Mining in Disaster Situations

Social media data (text) posted by individuals during disaster situations will be the input to this research work. It includes features such as identity, time, and location of the post. These inputs are limited to textual posts shared in social media such as Facebook and Twitter of English language.

Visualization of present accurate information about the disaster in categories such as content building situational awareness, factual subjectivity, action-oriented, supporting decision-making and contribute to an emotion-oriented segment.

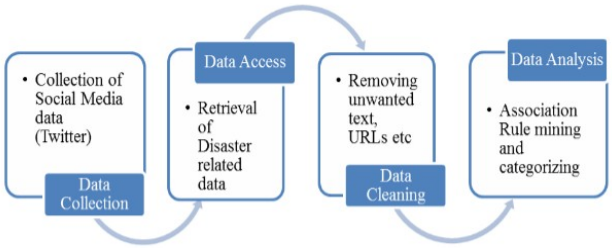

Figure 1: Approach to analyze data

Major components of research havebeen designed as data collection, data preprocessing, tokenizing, mining association rules for categories of decision making and finally detecting the relevant category when new data is entered. Text processing up to association rule generation has been doneover RapidMiner,and the rest of categorization is done using a software module designed with .Net technologies.

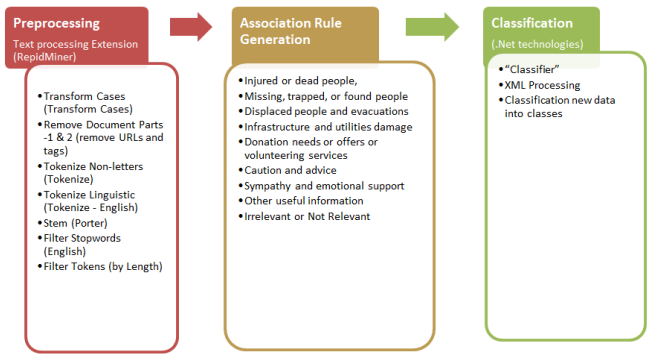

Figure 2: Design of Proposed System

A. *Text Processing*

Text mining (also referred to as text data mining or knowledge discovery from textual databases), refers to the process of discovering interesting and non-trivial knowledge from text documents. The common practice in text mining is the analysis of the information extracted through text processing to form new facts and new hypotheses, that can be explored further with other data mining algorithms. Text mining applications typically deal with large and complex data sets of textual documents that contain a significant amount of irrelevant and noisy information. Feature selection aims to remove this irrelevant and noisy information by focusing only on relevant and informative data for use in text mining. Some of the topics within text mining include feature extraction, text categorization, clustering, trends analysis, association mining,and visualization.

B. *Classifier - .Net Technology based tool*

To classify data into meaningful categories, a tool is developed byusing.Net technologies. It has been developed using C# in Visual Studio Environment.

Classifier Design

Input

1. Set of Association rules belong to eight classes (in XML format)
2. Incoming twitter data

Process

Classifying twitter data into nine classes

(8 defined classes and1 irrelevant class)

Output

Twitter messages categorized intonine classes

V. Implementation

With the availability of crisis-related posts collected from Twitter, human-labeled tweets, dictionaries of out-of-vocabulary (OOV) words, word2vec embeddings, and otherrelated tools[2] we were able to extract real-world data set consisting of tweets from Twitter. The resource consisting ofhuman-labeled data annotated by paired workers, annotated by volunteers, Word2vec embeddings trained using crisis-related tweets and Out-Of-Vocabulary (OOV) words and their meanings. Further actual disaster-related data available from 19 crises from 2013 to 2015 categorized into crisis types with the countries.

Two datasets were consisting of 1,236,610 and 5,259,681 tweets in English collected during 2014-09-06 to 2014-09-06: Pakistan Floods 2014 and 2014-08-10 to 2014-09-03: India Floods 2014. Datasets were consisting of tweet-ids, user-ids only, therefore, Imran, Mitra & all have published a tool to download full tweets content from Twitter. [23]

Annotation Scheme used in this work has been taken from the previous research published by Muhammad Imaran and team in their paper titled Twitter as a Lifeline: Human-annotated TwitterCorpora for NLP of crisis-related

2http://crisisnlp.qcri.org/

messages. They have taken these annotation schemes using input taken from formal crisis response agencies such as the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA).

Categorizing messages by informationtypes into 9 categories

a) **Injured or dead people**: Reports of casualties and/or injured people due to the crisis

b) **Missing, trapped, or found people**: Reports and questions about missing or found people

c) **Displaced people and evacuations**: People who have relocated due to the crisis, even for a short time (includes evacuations)

d) **Infrastructure and utilities damage**: Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored

e) **Donation needs or offers or volunteering services**: Reports of urgent needs or donations of shelter and supplies such as food, water, clothing, money, medical supplies or blood; and volunteering services

f) **Caution and advice**: Reports of warnings issued or lifted, guidance and tips

g) **Sympathy and emotional support**: Prayers, thoughts, and emotional support

h) **Other useful information**: Other useful information that helps understand the situation

i) **Not related or irrelevant**: Unrelated to the situation or irrelevant

The nine category types (including two catch-all classes: "Other Useful Information" and "Irrelevant") used by the UN OCHA

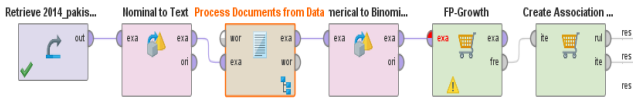

Figure 3: Operators for Process Documents from Files operator

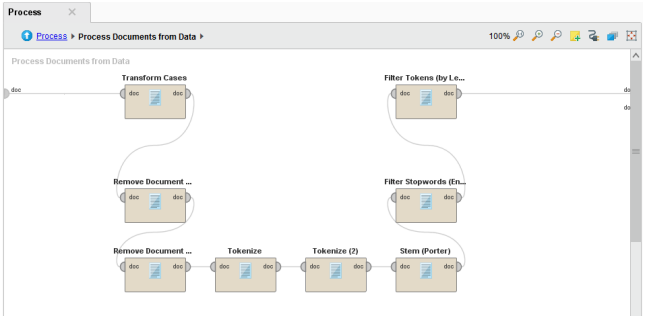

Figure 4: Operators within the process documents from files nested operator

The final result is the set of association rules. In Table View, table grid presents the generated association rules, with one rule in each row. For example, the first row states "IF Kashmir THEN flood" with a Support level of 0.321 and Confidence level of 1.00. This rule means that in 32 of the 100 documents, words with stem Kashmir and flood appear together. Furthermore, in 100% of the documents where a word derived from the stem Kashmir appears, at least one word derived from the stem flood is observed.

A. Clustering tweets using Association Rules

After generating association rules for each category, as the second module of this research, a tool is created to classify tweets automatically using Association Rules for each category name "Classifier" using.Net technologies. The development language used is C# and to input association rules to the classifier XML format has been used. Using the write operator in RapidMiner all the association rules generated was stored in an XML.

```
<object-stream>
  <AssociationRules id="1" serialization="custom">
    <com.rapidminer.operator.AbstractIOObject>
      <default>
        <source>Create Association Rules</source>
      </default>
    </com.rapidminer.operator.AbstractIOObject>
    <com.rapidminer.operator.ResultObjectAdapter>
      <default>
        <annotations id="2">
          <keyValueMap id="3"/>
        </annotations>
      </default>
    </com.rapidminer.operator.ResultObjectAdapter>
    <AssociationRules>
      <default>
        <associationRules id="4">
          <com.rapidminer.operator.learner.associations.AssociationRule id="5">
            <confidence>0.9629629629629629</confidence>
            <totalSupport>0.4642857142857143</totalSupport>
            <lift>0.9804713804713805</lift>
            <laplace>0.9879518072289157</laplace>
            <gain>-0.5</gain>
```

Figure 5: XML file of Association Rules

An algorithm is written to retrieve association rules of all nine categories and then to remove common rules for all categories since those rules cannot be used for classification. To uniquely identify the category of a new incoming tweet, the tweet is processed to remove stop words after tokenization and then stemming is done with porter stemmer. Then the key words of the tweet compared with all the association rules in all nine categories and the rule having highest confidence out of all matching rules identified and the category of that rule is considered as the category of a new tweet.

Any tweet not matching with any of the nine classes will be categorized as irrelevant for the subjective disaster. Such data is not considered,and once the data has been automatically categorized into nine classes, those subsets of data can be used to relief related or other humanitarian efforts related to the disaster.

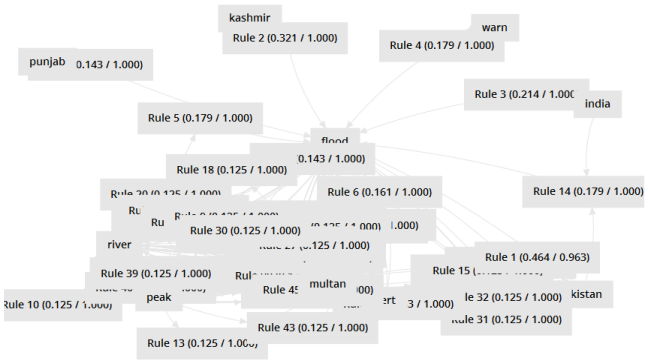

Figure 6: Graph view of Association Rules

```
Association Rules
[pakistan] --> [flood] (confidence: 0.963)
[kashmir] --> [flood] (confidence: 1.000)
[india] --> [flood] (confidence: 1.000)
[warn] --> [flood] (confidence: 1.000)
[river] --> [flood] (confidence: 1.000)
[alert] --> [flood] (confidence: 1.000)
[punjab] --> [flood] (confidence: 1.000)
[multan] --> [flood] (confidence: 1.000)
[peak] --> [flood] (confidence: 1.000)
[peak] --> [river] (confidence: 1.000)
[multan] --> [alert] (confidence: 1.000)
[peak] --> [alert] (confidence: 1.000)
[peak] --> [multan] (confidence: 1.000)
[pakistan, india] --> [flood] (confidence: 1.000)
[pakistan, alert] --> [flood] (confidence: 1.000)
[pakistan, multan] --> [flood] (confidence: 1.000)
[river, alert] --> [flood] (confidence: 1.000)
[river, multan] --> [flood] (confidence: 1.000)
[peak] --> [flood, river] (confidence: 1.000)
[flood, peak] --> [river] (confidence: 1.000)
[river, peak] --> [flood] (confidence: 1.000)
[multan] --> [flood, alert] (confidence: 1.000)
[flood, multan] --> [alert] (confidence: 1.000)
[alert, multan] --> [flood] (confidence: 1.000)
[peak] --> [flood, alert] (confidence: 1.000)
[flood, peak] --> [alert] (confidence: 1.000)
[alert, peak] --> [flood] (confidence: 1.000)
[peak] --> [flood, multan] (confidence: 1.000)
[flood, peak] --> [multan] (confidence: 1.000)
```

Figure 7: Association Rules of Caution and Advice Category

B. Summary of Evaluation

Following table shows the comparison of tweet categorization to nine pre-defined classes by two standard classifiers as well as the new association rule-based classifier introduced in this work. Final output shows that the classifier developed in this work classifies tweets into some classes with higher accuracy and others with an average level of accuracy. Therefore, the algorithm must be reviewed for optimization as a future work of this study.

Category	K-NN	Naïve Bayes	*Classifier
Injured or dead people	62.32%	68.22%	95.85%
Missing, trapped, or found people	50.00%	17.82%	62.86%
Displaced people and evacuations	53.33%	29.45%	70.42%
Infrastructure and utilities damage	62.50%	43.94%	26.97%
Donation needs or offers or volunteering services	64.00%	55.05%	72.11%
Caution and advice	55.00%	34.00%	27.45%
Sympathy and emotional support	70.00%	53.45%	91.53%
Other useful information	55.38%	54.36%	82.88%
Not related or irrelevant:	0.00%	4.22%	92.59%

Figure 8: Accuracy of different classifiers

VI. CONCLUSION

Social media data can be taken for numerous activities,and in this research work, it is given focus to text mining during natural disasters to assist decision making for relief authorities. With the massive use of social media during disaster situations incoming information filtering is very much important for situational awareness

Data categorization is done automatically using a tool developed, named "Classifier" and in this tool, a novel approach is used other than traditional classification algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest (RF),etc. which were used in related work of literature.

The algorithm is written with a novel approach using association rules already generated using human – annotated disaster dataset and searching the best category by giving priority to the highest confidence of matching rules for key words of new data.

Since we are using a real dataset consisting with tweets posted during natural disasters such as Pakistan Flood 2014 we had to preprocess a data to remove unnecessary words, URLs, tags etc. and text processing techniques were used using RapidMiner Text Processing extension. And then to generate association rules, FP- Growth algorithm is used.

The focusis given in this work to text only of tweets though they contain lot more features such as images, videos, geo-tags, user-tags, etc. In this work text posted in a tweet is considered and preprocessing is done to extract key words. Therefore, not like other text mining related researches, we ended up with a very small set of words to generate patterns.

And this approach is designed to automatically classify tweets into informative classes in a known disaster in a known location. By removing geographical keywords, we can apply the same model to Sri Lanka during flood situations because disaster-related keywords will remain the same since we have considered confidence of 90% for English words.

REFERENCES

[1] J. A. Obar and S. S. Wildman, "Social Media Definition and the Governance Challenge: An Introduction to the Special Issue by Jonathan A. Obar, Steven S. Wildman :: SSRN."

[2] P. Kallas, "Top 15 Most Popular Social Networking Sites and Apps [August 2017]," *DreamGrow*, 02-Aug-2017. [Online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/. [Accessed: 14-Aug-2017].

[3] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Practical extraction of disaster-relevant information from social media," 2013, pp. 1021–1024.

[4] S. Anson, H. Watson, K. Wadhwa, and K. Metz, "Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors," *Int. J. Disaster Risk Reduct.*, vol. 21, pp. 131–139, Mar. 2017.

[5] N. Pandey and S. Natarajan, "How social media can contribute during disaster events? Case study of Chennai floods 2015," in *2016 International Conference on*

Advances in Computing, Communications,and Informatics (ICACCI), 2016, pp. 1352–1356.
[6] J. D. Fraustino, B. Liu, and Y. Jin, "Social media use during disasters: a review of the knowledge base and gaps.," 2012.
[7] Y. Kryvasheyeu *et al.*, "Rapid assessment of disaster damage using social media activity," *Sci. Adv.*, vol. 2, no. 3, pp. e1500779–e1500779, Mar. 2016.
[8] Z. Li, C. Wang, C. T. Emrich, and D. Guo, "A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods," *Cartogr. Geogr. Inf. Sci.*, vol. 0, no. 0, pp. 1–14, Feb. 2017.
[9] "Importance of Disaster Management." [Online]. Available: https://targetstudy.com/articles/importance-of-disaster-management.html. [Accessed: 08-Jul-2017].
[10] "What is social media? - Definition from WhatIs.com." [Online]. Available: http://whatis.techtarget.com/definition/social-media. [Accessed: 08-Jul-2017].
[11] A. Kaur and D. Chopra, "Comparison of text mining tools," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2016, pp. 186–192.
[12] M. Kibanov, G. Stumme, I. Amin, and J. G. Lee, "Mining Social Media to Inform Peatland Fire and Haze Disaster Management," *Soc. Netw. Anal. Min.*, vol. 7, no. 1, Dec. 2017.
[13] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing Social Media Messages in Mass Emergency: A Survey," *ACM Comput. Surv.*, vol. 47, no. 4, pp. 1–38, Jun. 2015.
[14] S. Vieweg, C. Castillo, and M. Imran, "Integrating social media communications into the rapid assessment of sudden onset disasters," in *International Conference on Social Informatics*, 2014, pp. 444–461.
[15] D. Velev and P. Zlateva, "Use of social media in natural disaster management," *Intl Proc Econ. Dev. Res.*, vol. 39, pp. 41–45, 2012.
[16] R. L. Briones, B. Kuch, B. F. Liu, and Y. Jin, "Keeping up with the digital age: How the American Red Cross uses social media to build relationships," *Public Relat. Rev.*, vol. 37, no. 1, pp. 37–43, Mar. 2011.
[17] A. Kao and S. R. Poteet, *Natural Language Processing and Text Mining*. Springer Science & Business Media, 2007.
[18] M. Rajman and R. Besançon, "Text Mining: Natural Language techniques and Text Mining applications," *SpringerLink*, pp. 50–64, 1998.
[19] E. Younis, *Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study*. 2015.
[20] "Sensing Social Media: A Range of Approaches for Sentiment Analysis | SpringerLink." [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-43639-5_6. [Accessed: 08-Jul-2017].
[21] F. Villarroel Ordenes, S. Ludwig, K. de Ruyter, D. Grewal, and M. Wetzels, "Unveiling What Is Written in the Stars: Analyzing Explicit, Implicit, and Discourse Patterns of Sentiment in Social Media," *J. Consum. Res.*, vol. 43, no. 6, pp. 875–894, Apr. 2017.
[22] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI Soc.*, vol. 30, no. 1, pp. 89–116, Feb. 2015.
[23] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages," *ArXiv Prepr. ArXiv160505894*, 2016.