TWITTER ANALYTICS FOR INTEGRATED RESEARCH IN BIODIVERSITY

ASIAN CONFERENCE ON REMOTE SENSING ACRS 2019

Sameer Saran (1), Laksh Singla (2), Priyanka Singh (1)

[1] Indian Institute of Remote Sensing, ISRO, 4 Kalidas Road, Dehradun, Uttarakhand, India.
[2] Birla Institute of Technology and Science Pilani, Vidya Vihar, Pilani, Rajasthan, India

Email: saran.iirs@gmail.com; lakshsingla@gmail.com; priyanka.iirs@gmail.com

KEY WORDS: Social Media, Twitter Analytics, Sentiment Analysis, Geo-Visualization, Biodiversity

ABSTRACT: With the exponential rate of growth in usage of social media, the amount of information generated by the social media users is substantially growing which requires more human resources to analyse everything, thus not to miss the valuable information. Nowadays, every Facebook post, every Tweet, every blog entry, every social media platform is generating a new bit of information that may produce situational assessment or awareness in various domains. While this can and is being done, the research carried out in response to this is the development of web based geo-visual interface to forage and filter the place, time and theme from Twitter, and thus provides overview and detail on geographical, temporal and thematic tweets for further analytics. The primary streams of this research are – (i) extraction of geographic information from geotagged tweets and geocoding of that extracted information, (ii) web-based geo-visualization for text artifacts foraging and sense making, and (iii) sentiment analysis by generating the sentiment time series that elicits strong positive, negative ad neutral sentiments from users. By considering the importance of Twitter as one of the premiere social media platforms, the application is executed on several themes and results are having important implications for social intelligence and analytics. This web based place-time-theme indexing application uses biodiversity domain as a part of Indian Bioresource Information Network (IBIN) project of India. The IBIN project is aimed to create and maintain a digitized collection of the biological resources of India garnished from the published information sources and serve it through a common web platform to a diverse range of end users. The diverse amount of biodiversity data is available on social media platform has exploded in the last decade, but making these data available in real-time for generating the useful insights and patterns requires a considerable investment of time and work, both vital considerations for organizations and institutions looking to validate the impact factors of these online works. Therefore, this research model may embrace passion and proactiveness for conservation and protection of bio-resources. This work interrogated the social media content as a relevant source of information and integrated various twitter contents with filtering and foraging mechanisms, derivation of data facets such as sentiments, and visual methods on map for schematizing analyses. This idea to harness and drive new insights and stories from social media messages will reduce the possibility of noisy data and focus only on the information relevant to the particular theme which may leads to different types of analytical insights.

1. INTRODUCTION

Due to rapid development in World Wide Web, the acceptance and penetration of social media leads to unrivalled growth by changing the way of communication and interaction among people. According to Howard (2011), "We use Facebook to schedule the protest, Twitter to coordinate, and YouTube to tell the word". Among social media platforms, Twitter is becoming the prominent source of discussion and express their opinions on any current topic and thus, providing a vast platform for sentiment analysis or opinion mining.

Twitter is a very popular and reliable microblogging platform that allow its users to share their personal opinions in form of tweets (short messages up to 140 characters) (Sarlan et al. 2015; Lai 2010; Lohmann et al. 2012). The Twitter users can discuss on diverse topics such as current national and international issues, politics, personal life events, etc (Rambocas & Gama 2013). However, the tweets with fewer characters are shared by using a short form of words and symbols followed by hashtag (#). Therefore, the tweets can be used to find analyse and predict the viewpoints and sentiments for any current topic. For example, Twitter data has already been used to predict box office revenues for movies (Asur & Huberman 2010), stock market prediction (Bollen et al. 2011), identify the clients with negative sentiments (Thet et al. 2010), etc. Twitter, with over 500 million users and million messages per day, has now became an valuable source for any organizations to analyse the sentiment of the tweets by the public about their services, products, and even about competitors (Saif et al. 2012). Jose et al. (2010) emphasized that the sentiments generated from social media posts in form of voluminous tweets with the mammoth growth of the world wide web, blogs, forums, reviews or any discussion groups are available for analysis, thus making the Internet fastest, comprising and easily accessible medium for sentiment analysis. Sentiment analysis is the study based on natural language processing technique of extracting and deriving the user's sentiment from the raw data, where raw data is the online text that is exchanged by users through social media (Tang et al. 2009). The conceptual approach of sentiment analysis is to identify the polarity of overall data on particular topic by assigning them positive, negative or neutral grades. Polarity refers to the classification of a text or sentence in positive or negative (Sharma & Dey 2012). Therefore, semantic analysis (opinion mining) is dedicated to discover “what others think”, “how positive (or negative) are people” and “what would people prefer most” from online data (Pang et al. 2002). Due to online availability of rich text data opens a gateway to new opportunities for scientific research on data mining and natural language processing (Liu et al. 2011a; 2011b; 2011c; Jiang et al. 2011).

2. TWITTER SENTIMENT ANALYSIS

For analysing the polarity of tweets or comments, lexicon and machine learning approach are used to identify and classify the polarity of text (Ding et al. 2008; Taboada et al. 2010; Taboada et al. 2011; Annett & Kondrak 2009).

Lexicon Approach

This approach uses the predefined list of words where each word is associated with a specific sentiment (Goncalves et al. 2013), varied as per the context in which they were created and are involved in calculating the semantic orientation of texts or phrases in the documents (Taboada et al. 2011). Besides this, Sharma and Dey (2012) also states that a lexicon method is used in “detecting word-carrying opinion in the corpus and then to predict opinion expressed in the text”. Annett & Kondrak (2009) has described the following basic paradigm for lexicon-based analysis:

i. Preprocessing of each tweet by removing punctuation
ii. Initialization of total polarity score (s) equal 0 -> s=0
iii. Check whether specific token is present or not in a dictionary, then If token is positive, s will be positive (+), and If token is negative, s will be negative (-)
iv. Analyze the total polarity score of tweet post: If s > threshold, classify tweet post as positive, and If s < threshold, then consider tweet post as negative

However, in case of emotions and short hand texts, the accuracy of lexicon method is reduced, as they are not the part of predefined sentiment lexicon or dictionary (Khan et al. 2015). Therefore, Hu et al. (2013) proposed a novel sentiment analysis which considers the short and emotional texts in a unified framework. The performance of this novel method does not shows stability on few emotional signals when used on datasets from different domains (and short hand texts).

Machine Learning Approach

The machine learning based sentiment analysis is often relied on supervised classification algorithm where sentiment detection is considered as a binary values to classify in positive and negative (Sharma and Dey 2012). This approach required labelled datasets (Goncalves et al. 2013) to train the sentiment classifiers using feature vectors such as unigrams or bigrams (Pang et al. 2002) for categorizing them into positive, negative and neutral sentiments. Annett & Kondrak (2009) demonstrated a basic paradigm for creating feature vector:

i. Apply a part of speech tagger to each tweet post
ii. Collect all the adjectives for entire tweet posts
iii. Make a popular word set composed of the top N adjectives
iv. Navigate all of the tweets in the experimental set to create the following:
    - Number of positive words
    - Number of negative words
    - Presence, absence or frequency of each word

In the case of strongly negative or positive values that reflects a mixed perspective, are correctly captured in the shifted value. Furthermore, Annett & Kondrak (2009) stated that machine learning methods generate a fixed number of the mostly happening popular words by assigning an integer value instead of the frequency of the word in the Twitter. Also, Goncalves et al. (2013) has illustrated the limitations of machine learning approach that proved to be more suitable for Twitter analysis than the lexicon-based method.

This paper tried to analyse the geo-tagged twitter posts on biodiversity keywords using machine learning approach in real-time. The biodiversity keywords tweeted using Twitter hashtags (e.g., #tiger, #conservation, #forestfire) are classified as positive, negative, and neutral tweets and are then plotted on Google map by extracting and geocoding the locations from tweets. A web interface tool was developed to aid in the twitter sentiment analysis task.

3. RELATED WORK

In past few years, sentiment analysis has grown exponentially in area of Natural Language Processing with research ranging from classification of opinions, messages and documents (Pang and Lee 2008) to learn the polarity exists in words and phrases (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006). Due to character limitations on Twitter, classification of the sentiments of tweets is most likely similar to the sentence-level sentiment analysis (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004). In last decade, there have been a number of research articles/papers on Twitter sentiment analysis (Jansen et al. 2009; Pak and Paroubek 2010; O'Connor et al. 2010; Tumasjan et al. 2010; Bifet and Frank 2010; Barbosa and Feng 2010) and many researchers have begun to investigate various ways of automatically collecting training data (Pak and Paroubek 2010; Bifet and Frank 2010). (Barbosa and Feng 2010) exploited the existing Twitter sentiment sites in seek of training data and used 1000 manually labelled tweets for training and another 1000 manually labelled tweets for testing purpose. (Davidov et al. 2010) used hashtags for preparing training data, but they limited their experiments to sentiment/non-sentiment classification. The training data of (Go et al. 2009) consisted of tweets with emotions like ":)" and ":(" and they marked these emotions as noisy labels, which is similar to (Read 2005). (Davidov et al. 2010) marked 50 tags and 15 smileys as noisy labels to identify and classify diverse sentiment types of tweets. Gamon (2004) performed sentiment analysis on feedback data of Global Support Services survey by analysing the role of linguistic features.

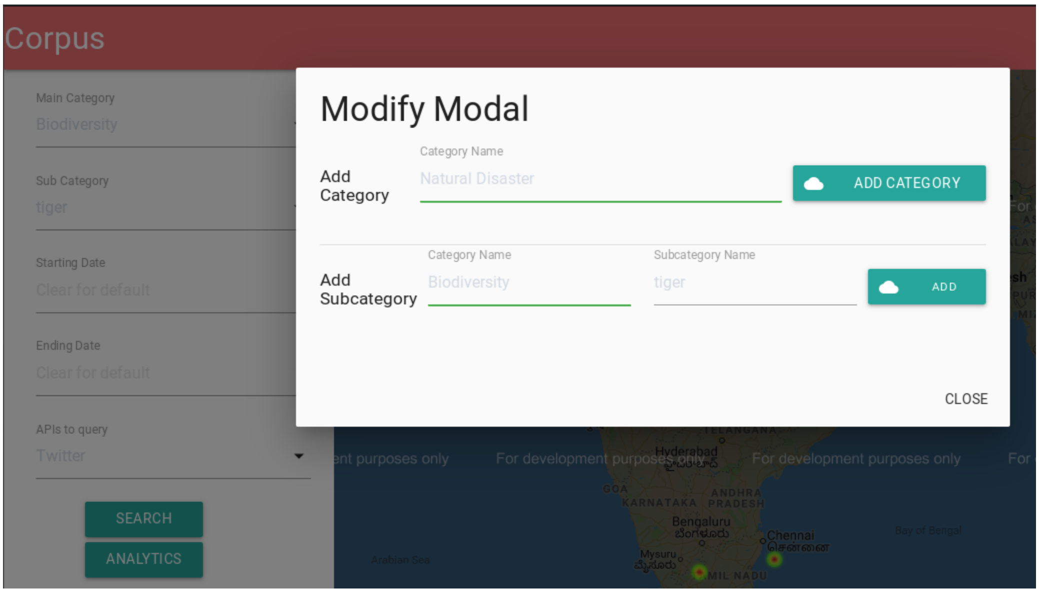
4. PROPOSED METHODOLOGY

The study presented in this paper is categorized into the following phases:-

a. Fetch the tweets from the Twitter

Tweepy is used to provide an abstraction between the tweets data and the application. The API class provides access to the entire twitter RESTful API methods. Each method can accept various parameters and return responses. Tweepy model class instance is returned on utilizing the most methods. Secure connection with the twitter client is made using OAuth2 method of authentication. Since at the current stage of development, we require only read only methods to public information without any user context, therefore it is just sufficient to pass consumer_token and consumer_secret variables to the API service. Following parameters are passed while querying the API:-

- Retweets are not allowed (-filter:retweets)
- Language is restricted only to english (lang='en')
- Location from where the tweet is made is restricted to a circular perimeter around India (geocode='22(lat), 78(long), 2000(radius)')



Main Category: Biodiversity

Sub Category: tiger

Starting Date: Clear for default

Ending Date: Clear for default

APIs to query: Twitter

Category Name: Natural Disaster

Category Name: Biodiversity

Subcategory Name: tiger

Figure 1 Screenshot to add keywords for twitter analysis

b. Preprocessing and cleaning the tweets

Since tweets contains lots of mentions and hashtags to different topics, the unnecessary punctuation is removed from the data. Since we have to geocode the data, and therefore, pass it through a Named Entity Recognizer, the stop words and lemmatization techniques are not applied since it would hinder in its functioning. Simple regular expression search and replace is sufficient for cleaning for sentiment analysis.

c. Sentiment analysis on given tweets

Sentiment analysis is mining of text which identifies and extracts subjective information in source material, and helps monitor the underlying theme of unstructured textual data. TextBlob is a Python library used in this experiment for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification and translation. The textblob's sentiments module contains two sentiment analysis implementations – PatternAnalyzer (based on the pattern library) and NaiveBayesAnalyzer (an NLTK classifier trained on a movie reviews corpus). The default implementation is PatternAnalyzer, but it is overridden by the NaiveBayesAnalyzer, provided in the same library itself.

The tweets are then classified into positive, neutral and negative, as depicted in **Figure 2**. The relative balance of positive over negative gives a general trend of the mood surrounding the given data. The NaiveBayes classification method will categorise it as a positive tweet despite capturing that it depicts the sad state of affairs in a particular region. Hence, it is an assumption that the given ratio gives a good estimate of the sentiment of the underlying tweets data.

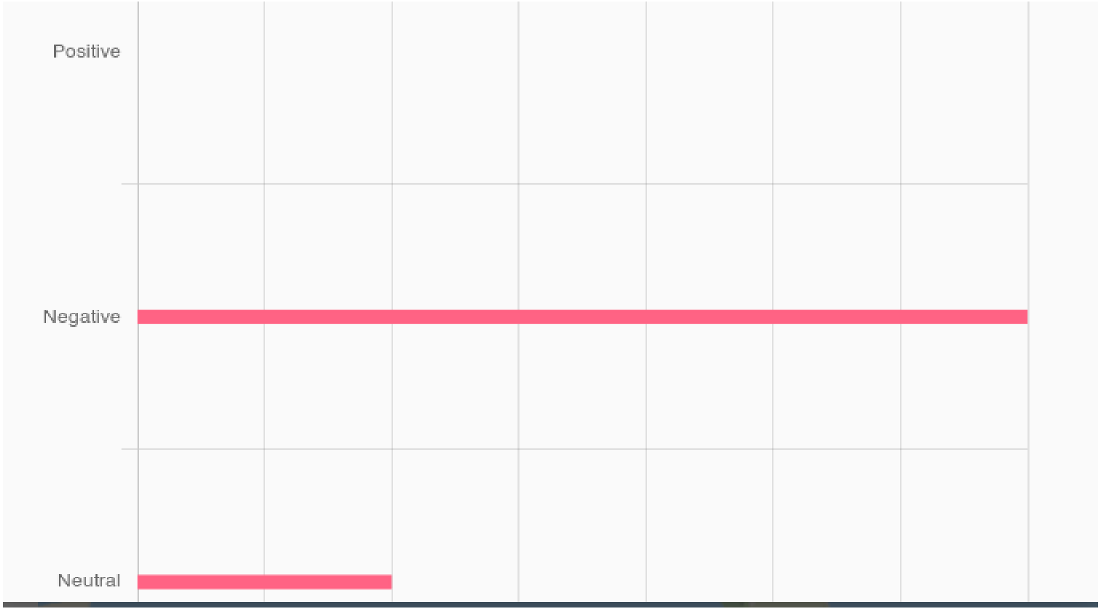


Figure 2 Sentiment analysis of tweets

d. Extract location names using Named Entity Recognition on the given tweet

Named Entity Recognition can identify individuals, companies, places, organizations, cities and other various types of entities. For this purpose, python’s spaCy package has been used, which has been trained on the OntoNotes5 corpus which is a large annotated corpus comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows) in three languages (English, Chinese, and Arabic) with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference). The reason for using spaCy rather than Stanford NER was the usability and the time taken to analyse the text. Since tweets are small pieces of text, therefore not much context is required and hence a simpler library is sufficient for recognition.The statistical models in spaCy are custom-designed and provide an exceptional performance mixture of both speed, as well as accuracy. The current architecture used has not been published yet, but it is a statistical model rather than a Deep Learning model, hence word embeddings etc. are not used in recognition, but the reduced accuracy is offset by the improved accuracy.

e. Geocode the given tweet

The tweet is geocoded by using Nominatim API. Nominatim is a tool to search OSM data by name and address (geocoding) and to generate synthetic addresses of OSM points (reverse geocoding). The search feature of the Nominatim API is primarily utilised in geocoding the data. The search API allows to look up a location from a textual description. The search query may also contain special phrases which are translated into specific OpenStreetMap (OSM) tags (e.g. Pub => amenity=pub). OpenStreetMap is both neutral and transparent. OSM data is initially imported using osm2pgsql. Nominatim uses its own data output style 'gazetteer', which differs from the output style created for map rendering. Although the implementation is hidden from the user, following are the steps followed by it:

- An address normalizer that takes an arbitrary string and normalizes it
- A geocoder that does some magical fuzzy matching for names where the core algorithm is the Levenshtein Distance.
- Some interpolation of the street segments at the end to guess where the house is

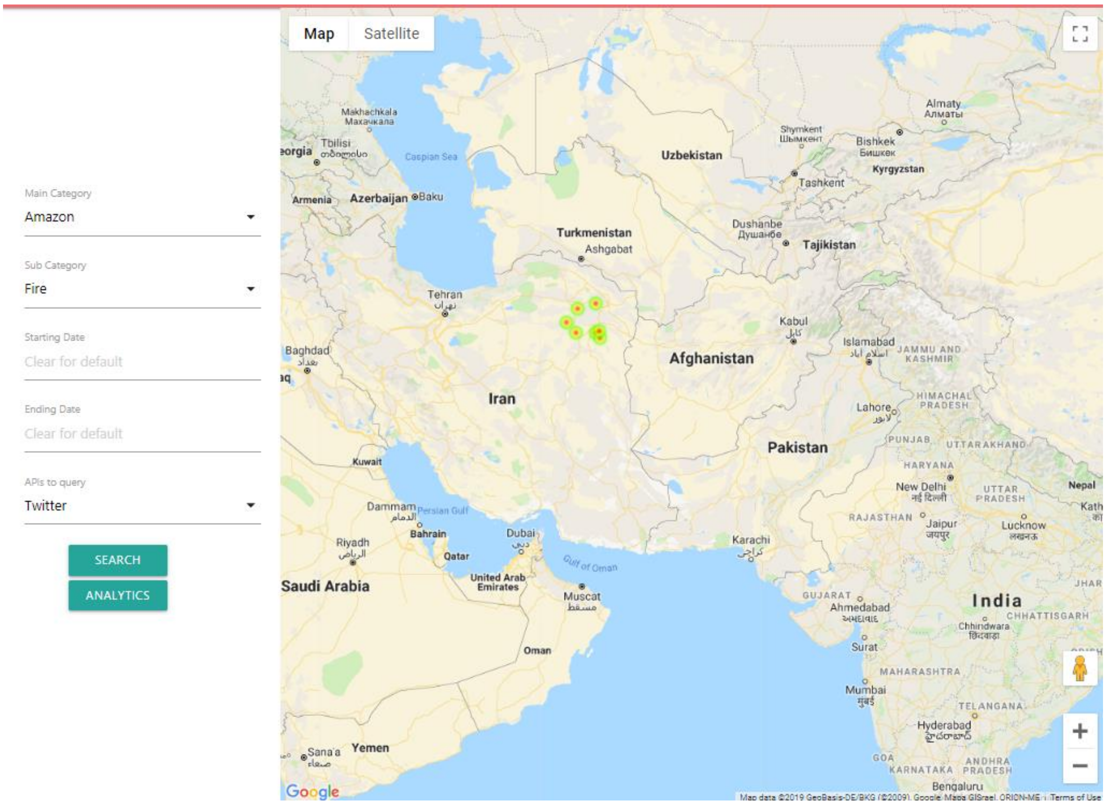
The Open Search API is updated approximately every 20 minutes and is mostly responsive, but it might experience downtime, in case there is huge data import.

f. Store the response in MongoDB database

MongoDB is a document database, which is used to store data in JavaScript Object Notation (JSON) format. It is a NoSQL database, which is easily scalable and can be parallelised. Although, it is unable to perform joins efficiently, but that feature is not required in unstructured data like tweets. The ability of MongoDB to generate GeoJSON file can be converted to shapefile for further geospatial analytical tasks. This methodology integrated Django with MongoDB to store the geo-tagged tweets in MongoDB as a record to perform tasks in future.

5. CASE STUDY

Using the keyword “Amazon fire” tested between August 23 and September 8, 2019.



Main Category
Amazon

Sub Category
Fire

Starting Date
Clear for default

Ending Date
Clear for default

APIs to query
Twitter

Figure 3 Output of twitter analytics

According to the plot shown in Figure 3, as Amazon forest fire is a current topic amongst the tweets, therefore a very high concentration of the tweets were geocoded to locations. After the analysing the public opinions through their tweets, set of 258 negative tweets, 739 neutral and 503 positive tweets are marked for such incident.

6. DISCUSSION

Twitter is an open social environment where users can express their opinions in form of tweet on different topics within 140-character limit. This poses a significant challenge to Twitter sentiment analysis since tweets data are often noisy and contain a large number of irregular words and non-English symbols and characters. In this paper, the public mood patterns are evaluated for the forest fire incidents of Amazon forest, as evidenced from a sentiment analysis of Twitter posts from August 23 to September 8, 2019. A set of 258 negative tweets, 739 neutral and 503 positive tweets are marked for such incident. This relates the fluctuations in macroscopic socio-economic indicators in the same time period. With such analysis, various attempts can be initiated to identify a quantifiable relationship between overall public mood and social, economic and other major events in the media and popular culture.

To conclude that sentiment analysis of minute text corpora (such as tweets) is efficiently obained via a syntactic, term-based approach that requires no training or machine learning. Sentiment analysis techniques rooted in machine learning yield accurate classification results when sufficiently large data is available for testing and training.

7. REFERENCES

Annett, M., & Kondrak, G. (2008). A comparison of sentiment analysis techniques: Polarizing movie blogs. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-540-68825-9_3

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010. https://doi.org/10.1109/WI-IAT.2010.63

Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference.

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-642-16184-1_1

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science. https://doi.org/10.1016/j.jocs.2010.12.007

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In WSDM'08 - Proceedings of the 2008 International Conference on Web Search and Data Mining. https://doi.org/10.1145/1341531.1341561

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th International Conference on Computational Linguistics. https://doi.org/10.3115/1220355.1220476

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. Processing.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. In COSN 2013 - Proceedings of the 2013 Conference on Online Social Networks. https://doi.org/10.1145/2512938.2512951

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. https://doi.org/10.3115/979617.979640

Howard, P. N. (2011). The Arab Spring's cascading effects. Pacific Standard, 23.

Hu, X., Tang, J., Gao, H., & Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology. https://doi.org/10.1002/asi.21149

Jose, A. K., Bhatia, N., & Krishna, S. (2010). Twitter sentiment analysis. In Seminar Report, National Institute of Technology Calicut.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification. In ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Khan A.R. (2015) Trade and Global Links. In: The Economy of Bangladesh. Palgrave Macmillan, London

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. https://doi.org/10.3115/1220355.1220555

Lai, P. (2010). Extracting Strong Sentiment Trends from Twitter. Nlpstanfordedu.

Liu, X., Li, K., Zhou, M., & Xiong, Z. (2011a). Collective semantic role labeling for tweets with clustering. In IJCAI International Joint Conference on Artificial Intelligence. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-307

Liu, X., Li, K., Zhou, M., & Xiong, Z. (2011b). Enhancing semantic role labeling for tweets using self-training. In Proceedings of the National Conference on Artificial Intelligence.

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011c). Recognizing Named Entities in tweets. In ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Lohmann, S., Burch, M., Schmauder, H., & Weiskopf, D. (2012). Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In Proceedings of the Workshop on Advanced Visual Interfaces AVI. https://doi.org/10.1145/2254556.2254701

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010.

Pang, B., & Lee, L. (2008). Presentation: Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

Rambocas, M., & Gama, J. (2013). Marketing Research : The Role of Sentiment Analysis. Working Papers (FEP) - Universidade Do Porto.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference.

Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-642-35176-1-32

Sarlan, A., Nadam, C., & Basri, S. (2015). Twitter sentiment analysis. In Conference Proceedings - 6th International Conference on Information Technology and Multimedia at UNITEN: Cultivating Creativity and Enabling Technology Through the Internet of Things, ICIMU 2014. https://doi.org/10.1109/ICIMU.2014.7066632

Sharma, A., & Dey, S. (2012). Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. International Journal of Computer Applications.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-basedmethods for sentiment analysis. Computational Linguistics. https://doi.org/10.1162/COLI_a_00049

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2009.02.063

Thet, T. T., Na, J. C., & Khoo, C. S. G. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science. https://doi.org/10.1177/0165551510388123

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.