Emotion Profiling, Trend Analysis and Empath Link Prediction to Government Responses using Twitter during the #IndiaFarmersProtest

Mohammad Talib
IIIT Delhi
Okhla Industrial Estate, Phase III
New Delhi, India
www.talibbinjawed@gmail.com

Kashish
MLNC,
Benito Juarez Marg,
South Campus, New Delhi, India
1111kashish1111@gmail.com

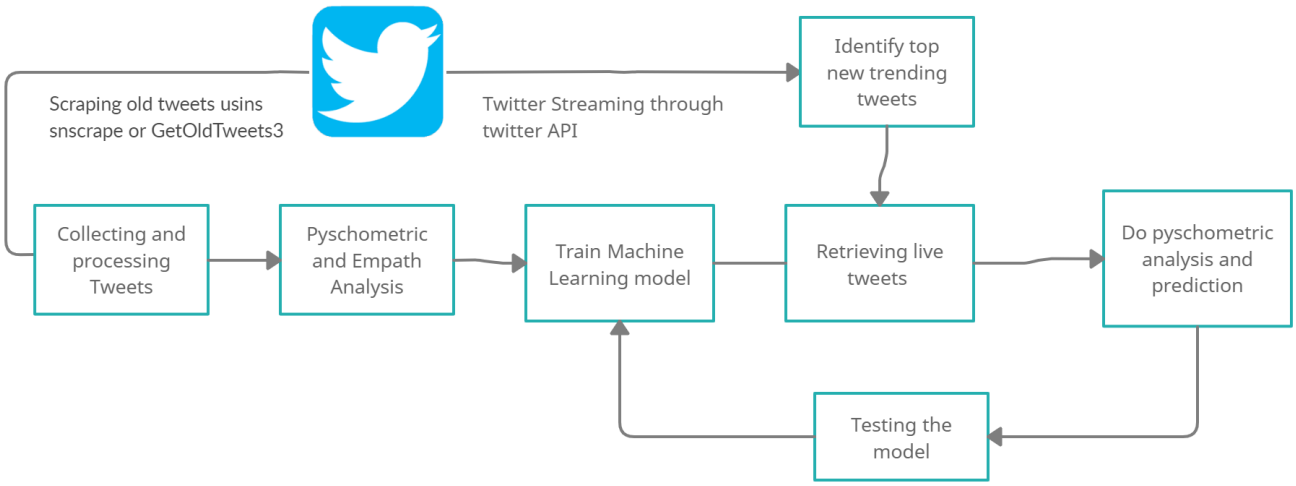

Figure 1. Graphical Abstract

ABSTRACT

Social media plays an important role in formulation of **collective and connective action,** and sentiments form a large shaping factor in these actions. These actions more often than not turn into protest which may or may not be very harmful for a state. We use twitter as open source data as input of the recent IndiaFarmersProtest as a framework for analysis. In this project we plan a quantitative and localised emotion and psychometric profiling by topic based vector space models for classification into empath categories and thus present a localized view of the opinion of the protest throughout India. We will use statistical tests to establish the significance of Government tweets or bulletins to have a statistical significance on the protest empathy. On positive results, moreover using Empath features/scores generated we plan to train machine learning models to predict change in sentiments in response to government response though news and social media and present an automated, 24x7 continuous system web app for analysing sentiment and forecasting civil unrest across India using open source indicators.

Keywords

Twitter Data Mining, Empath Analysis, Topic Signaling, Link/Empath Prediction.

1. INTRODUCTION

There is a rise of a personalized digitally networked politics in which diverse individuals address the common problems of our times such as economic fairness and climate change. Large-scale connective action is coordinated using inclusive discourses easily through social media. In the real world people's opinions are

based fundamentally in the virtual domain for most human operations on some particular basis and they can be crucially instigated on the actions of a large number of users. We have seen a decision essential be taken, just because a large no say something, which sometimes can cause harm to the society.

The dramatic increase in the number of users on social media platforms leads to the generation of huge amounts of unstructured text. Besides the exchange of information, social media is a remarkably convenient medium to express the ideas and opinions. This can be seen in many cases such as the **Arab Spring** incidents and even the delhigangrape protests in which social media played an important role for organisation of protests.[1]

1.1 #IndianFarmersProtests

These protests may or may not be unproductive and it is important to keep in mind the rights of the citizens while maintaining security. In such cases, social media analysis can play a huge role in gauging the situation and making predictions and thus is very important. One such recent protest in India is the IndianFarmersProtest. The protest by thousands of farmers at Delhi's borders has focused on a range of agriculture issues in India with opposing views where the government says the laws will bring necessary modernization and private competition to an ailing sector while farmers say the laws were passed without consultation and will allow private corporations to control the prices of crops. Farmer unions and their supporters have demanded revocation of all the farm acts and they also want creation of a legally minimum support price for their crops and vegetation to ensure that corporates cannot control the prices. Many farmers and their union described the bill as "anti-farmer laws".

The protest principally ran peacefully for months but is however unpredictable and may turn out to be unrest as seen on 26th January of the protest. Moreover the protest has been guided by empathy of the farmers, 'anger', 'fear'. 'hope', 'distrust'. Analysing such sentiments could be key to assessing it and a method to gauge the reaction by government officials and linking government statements to the reactions of the protesters. Such that inflammatory statements and escalation of situations do not happen.

1.2 Motivation

We are principally moved by the protest and want to Analyze tweets related to the farmer's protest (something which the research community has been mostly silent on) and try to mathematically establish the empathy of the people in various regions and to predict the further activity that can be a cause of concern.

The social media platforms, such as Twitter, generate huge amounts of the text containing political insights, which can be mined to analyze the people's opinion and predict the future trends More often than not the sentiments in online social media form a basis for a collective action which mobilizes and gets translated into a form of planned protest and unrest.

Through data analysis of open source data like twitter one would be able to find out what the perception of the common public tends towards unrest, we would be able to provide information like localized sentiments in the form of mathematical data to better make policies. This would help the government and authorities assess the feelings of people and act on it.

This can also tell the possibility of a future protest and with how much intensity the protest in running. Moreover, we should be able to predict when such a protest may take the face of a social unrest and furthermore see how their responses are affecting and interacting with the protest they would be better able to manage the situation controlling the narrative. This may prevent much damage and save many lives and is of great interest to us.

1.3 Previous Work

There is a vast and much prevalent work on sentiment analysis such as [1][2][5][7], and recent state-of-the-art models such as **XLNet, BERT,** and **RoBERTa** based sentiment analysis models**.** However, we're especially impressed and interested in sentiment analysis by topic modeling methods such as **Empath,** which we discuss in detail further in our methodologies.

Work on unrest prediction has been done using **Hidden Markov Models (HMMs) on GDELT[9]** (Global Data on Events, Location, and Tone), **Support Vector Machine on LWIC topics[6]** based clustering and **context graph based neural networks** to make probabilistic predictions among others. A variety of work has been done on predicting social events based on open source data[9-14]. **EMBERS** is a continuous unrest monitoring system for 11 Latin Countries using **Dynamic Query Expansion and Cascades models.**

Sentiment analysis stands as the prominent research topic in demand under the **Natural Language Processing (NLP**). The fundamental objective of this research topic is to spot out the emotions and opinions of the customers or users via a text basis. Even though numerous research works have been carried put in this field through diverse models, sentiment analysis is still considered a challenging problem with so many conflicts to be solved. Some of the existing challenges are due to the slang words, new accents, grammatical and spelling mistakes etc. This paper plans to make a literature review using different machine learning algorithms with various data[16][17].

Sentiment analysis techniques are increasingly used to grasp reactions from social media users to unexpected and potentially stressful social events. The paper[18][19][20] argues that, alongside assessments of the affective valence of social media content as negative or positive, there is a need for a deeper understanding of the context in which reactions are expressed and the specific functions that users' emotional states may reflect. To demonstrate this, we present a qualitative analysis of affective expressions on Twitter collected in Germany during the 2011 EHEC food contamination incident based on a coding scheme developed from Skinner et al.'s (2003) coping classification framework. Beyond the positive or negative tone, some people perceived the outbreak as a threat while others as a challenge to cope with.

1.4 Research Objectives/Questions

1.4.1 Statistical Data Analysis

- R1: What was the trend of Twitter usage in the protest? What were the key themes of tweets during the protest days?
- R2: What were the most frequently linked media resources in the protest? a) overall b) on 26th January
- R3: Who were the most active individual Twitter users/stakeholders during the protest?

1.4.2 Quantitative Psychometric Analysis

- R1: What was the emotional classification or sentiment of Twitter in the protest?
- R2: Did all states show similar emotional sentiment? Which were the states with most positive and most negative emotional sentiments
- R3: How did the sentiment change over time?

1.4.3 Empath Based Prediction

- R1: Did government responses affect emotion sentiment of the data? How did it affect the sentiment?
- R2: Can we link sentiment changes to specific government responses?
- R3: Can we predict future sentiment of the protest to a certain government response?

In our final output we present an automated, 24x7 continuous system web app for forecasting civil unrest across India using open source indicators as of now just tweet and news sources.

2. MATERIALS AND METHODS

2.1 Data

2.1.1 Twitter Data Collection

TowardsDatascience describes various methods to mine and scrape twitter data. We plan to collect and curate our own dataset using Twitter's official **Tweepy API** for current tweets and scrape older tweets using **snscrape** till **GetOldTweets3** gets fixed. The pipeline for all 3 is referenced below in references. [3]

2.1.2 QuerySet

The data will be collected using 3 separate approaches in parallel – Content Based Query, Location Based Query and User Based Query..

Content-based Query: To collect the relevant Twitter data, will explore the trending and most popular hashtags for each of the Indian states and manually curate the list of hashtags related to the Farmers Protest. We also go through historical Farmer Protest related tweets manually to find and subsequently mine the most popular hashtags related to the same, which may not be trending. Further, in order to automate the collection of relevant tweets, we will formulate generic queries like 'farmers', 'farmers protest' etc. and collect the state-wise protest related twitter data using the same. This approach focuses on getting all tweets which are talking about Multiple queries will be built by joining terms related to the protest with the name or common aliases of the region. Some of the terms used include - 'farmers', 'kisan', 'kisanekta', 'lockdown', etc. In addition, popular hashtags like #farmersprotest, #farmers, #kisanekta, #indianfarmersprotest These will be used in combination of localized context with the name or popular alias of India or various states. Examples of a popular alias are 'Orissa' for Odisha, 'TN' for Tamil Nadu, 'UP' for Uttar Pradesh, or spelling mistakes like 'chatisgarh' for Chhattisgarh.

Location-based Query: Tweets are collected for globally trending Farmers Protest hashtags related hashtags, particularly #farmersprotest, #farmers, #kisanekta, #indianfarmersprotest, and then filtering tweets based on the User Location. We will also create a list of location filters for various states. These are the state names, aliases (as explained above) and names of popular cities in those states. Using these, we first filter out all tweets having 'india' in their user location, and then sort them based on keyword matches of tokens in the user location with the above list. The User Location was lowercased before matching.

User-based Query: Tweets are collected for globally associated users with Farmers Protest, like Narendra Modi, PMO India, Meena Harris, Kisan Ekta Morcha etc, then filtering tweets based on the keywords like farmers, protest, etc. We will also create a list of keyword based filters periodically.

We plan to have a minimum of 35,000 tweets at the least. with the maximum ranging in lakhs.

2.1.3 Preprocessing

We followed the below steps to pre-process the data and reduce the noise:

1. Lowercasing
2. Tokenization
3. Links Removal
4. Translation of non-english text: Using Google Translate API
5. Removal of all other non alphanumeric characters, allowing only standard English alphabets
6. Stop Words Removal: Performed Stop Words removal, using NLTK, for English
7. Lemmatisation: NLTK Wordnet Lemmatizer or spaCy

2.2 Methodology

2.2.1 Dataset Statistical Analysis

We do statistical analysis of the whole dataset to determine the statistics, content and themes of the protest. This is to answer basic questions such as what were the trend and characteristics, who were the stakeholders and active users and what content was shared. We do so by common computational exploratory processes and provide simple answers. Many such simple data analysis are highly beneficial such as xyz

2.2.2 Quantitative Psychometric Analysis

We will quantitative psychometric analysis of all tweets by associating them with categorical emotions. For such categorization we use topic signaling as in Empath. [4]

Empath is a tool for analyzing text across lexical categories (similar to LIWC), and also generating new lexical categories. It uses the skip-gram networks to construct a vector space model (VSM). VSMs allow Empath to discover member terms for categories. Empath uses deep learning to learn a neural embedding across more than 1.8 billion words of modern fiction.

Given a small set of seed words that characterize a category, Empath uses its neural embedding to discover new related terms. Empath also analyzes text across some inbuilt categories, including emotions, which are then used to identify the emotion associated with a text in our case.

2.2.3 *Empath Based Prediction*

First we use statistical tests to establish the significance of Government tweets or bulletins to have a statistical significance on the protest empathy. Such work has recently been done on a large scale for coronavirus trend predictions. One such work we take as our base is Baani et. al. as referenced below. [5]

Then we plan to either use statistical features of Empath scores generated to train machine learning prediction models, or either use vector space models to train regression of emotional sentiment values. We plan to deploy a similar system/dashboard similar to EMBER in the indian context.

3. RESULT AND EVALUATION

3.1 Results

3.1.1 *Dataset Statistical Analysis*

O1: Most commonly used words and such trends as a wordcloud

O2: Various graphs, barcharts and tables detailing the statistics of the dataset and their characteristics

3.1.2 *Quantitative Psychometric Analysis*

O1: Psychometric and empath analysis of the dataset over time periods. Radar plots of various emotions.

O2: Empath analysis to particular to Indian states. Radar plots of various emotions.

O3: Time-series analysis of Emotions in the protest over time as graphs

3.1.3 *Empath Based Predictions*

O1: Statistical significance of Official Government bulletins and tweets on different sentiments.

O2: Predictive models for text on further change in emotional sentiments of the protest.

3.2 Evaluation

All outputs need no evaluation except sentiment prediction models in Deliverable 3.O2. The evaluation criteria will accordingly be with respect to the model decided. Likely Mean Square Error of the Empath scores or classification accuracies.

4. DISCUSSION

4.1 Limitations

The Empath models reports high Pearson correlations vs LIWC's categories, it is possible that other more qualitative properties are important to lexical categories. Two lexicons can be statistically similar on the basis of word counts, and yet one might be easier to interpret than the other, offer more representative words, or present fewer false positives or negatives.

At a higher level, the number and kinds of emotions available in Empath though more than sentiment analysis present are predefined. They may not offer the right balance and breadth of emotions. Further, choice of seed words can be important. The given text in our data may not be enough to determine a category.

Finally, Empath is trained on a predefined fiction based dataset provides a powerful model for generating lexical categories, but our wee are more likely related to news. It is arguable other datasets may perform better at specifying categories. News as a training dataset might give it a more nuanced view of politics and protests. We see potential for training Empath on other texts beyond fiction.

4.2 Further Scope

Such psychometric analysis techniques can be used for predicting unrest using Online Social Media and Empath and Topic Classification. Such models already exist in literature such as Rostyslav Korolov et. al. as referenced below, however can be improved using time series modeling.

5. ACKNOWLEDGMENTS

Our thanks to Dr. Rajiv Ratn Shah for providing us with this golden opportunity to undertake this project and for this constant support and motivation in the course.

6. REFERENCES

[1] M. Patil and H. K. Chavan, "Event Based Sentiment Analysis of Twitter Data," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2018, pp. 1050-1054, doi: 10.1109/ICCMC.2018.8487531.

[2] Ahmed, Saifuddin and Jaidka, Kokil, Protests against #delhigangrape on Twitter: Analyzing India's Arab Spring (November 21, 2013). eJournal of eDemocracy and Open Government, 5(1), 28-58, 2013, Available at SSRN: https://ssrn.com/abstract=2413936

[3] Beck, Marthin. How to Scrape Tweets With snscrape. Medium - Better Programming. Available at https://medium.com/better-programming/how-to-scrape-tweets-with-snscrape-90124ed006af

[4] Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 4647–4657. DOI: https://doi.org/10.1145/2858036.2858535

[5] Jolly, Baani Leen Kaur & Aggrawal, Palash & Gulati, Amogh & Sethi, Amarjit & Kumaraguru, Ponnurangam & Sethi, Tavpritesh. (2020). Psychometric Analysis and Coupling of Emotions Between State Bulletins and Twitter in India during COVID-19 Infodemic. Available at arxiv: https://arxiv.org/abs/2005.05513

[6] Rostyslav Korolov, Di Lu, Jingjing Wang, Guangyu Zhou, Claire Bonial, Clare Voss, Lance Kaplan, William Wallace, Jiawei Han, and Heng Ji. 2016. On predicting social unrest using social media. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '16). IEEE Press, 89–95. https://dl.acm.org/doi/10.5555/3192424.3192441

[7] Rezapour, Rezvaneh-Shadi. (2018). Using Linguistic Cues for Analyzing Social Movements. Available at arxiv: https://arxiv.org/abs/1808.01742

[8] Eom Y-H, Puliga M, Smailović J, Mozetič I, Caldarelli G (2015) Twitter-Based Analysis of the Dynamics of Collective Attention to Political Parties. PLoS ONE 10(7): e0131184. https://doi.org/10.1371/journal.pone.0131184

[9] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, "Predicting social unrest events with hidden Markov models using GDELT," Discrete Dynamics in Nature and Society, vol. 2017, Article ID 8180272, 13 pages, 2017.

[10] N. Ramakrishnan, P. Butler, S. Muthiah et al., "Beating the news' with embers: forecasting civil unrest using open source indicators," in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1799–1808, ACM, New York, NY, USA, August 2014.

[11] S. Deng, H. Rangwala, and Y. Ning, "Learning dynamic context graphs for predicting social events," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1007–1016, ACM, Anchorage, AK, USA, August 2019.

[12] K. Nathan, "Predicting crowd behavior with big public data," in Proceedings of the 23rd International Conference on World Wide Web, pp. 625–630, ACM, Seoul, Korea, April 2014.

[13] W. Yang, X. Liu, J. Liu, and X. Cui, "Prediction of collective actions using deep neural network and species competition model on social media," World Wide Web, vol. 22, no. 1, 2018.

[14] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Spatiotemporal event forecasting in social media," in Proceedings of the SIAM International Conference on Data Mining, vol. 15, pp. 963–971, SIAM, Vancouver, Canada, April 2015.

[15] S. Ranganath, F. Morstatter, X. Hu, J. Tang, S. Wang, and H. Liu, "Predicting online protest participation of social media users," in Proceedings of the AAAI, pp. 208–214, Phoenix, AZ, USA, February 2016.

[16] G. Dharani Devi and Dr. S .Kamalakkannan. (2020). Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications. *International Journal of Advanced Science and Technology*, *29*(7), 1462 - 1471. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/15651

[17] Aisopos F, Papadakis G, Tserpes K, Varvarigou T (2012) Content vs. context for sentiment analysis: a comparative analysis over microblogs. In Proceedings of the 23rd ACM conference on Hypertext and social media, pp. 187–196. ACM

[18] Gaspar R, Pedro C, Panagiotopoulos P, Seibt B (2016) Beyond positive or negative: qualitative sentiment analysis of social media reactions to unexpected stressful events. Comput Hum Behav 56:179–191

[19] Ramteke J, Shah S, Godhia D, Shaikh A. Protest Election result prediction using Twitter sentiment analysis. In: 2016 international conference on inventive computation technologies (ICICT), vol. 1. 2016. p.

[20] Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 347–354. Association for Computational Linguistics

---xxx---