

Classification of Comments Supporting the Farmer Protest

Ajay Biswas
Supervisor: Dr. Tapas Kumar Mishra

October 28, 2021



Contents

- 1 Introduction
- 2 Motivation
- 3 Objectives
- 4 Literature survey
- 5 Methodology
- 6 Conclusion

Indian farm reforms 2020

- Farmers' Produce Trade and Commerce (Promotion and Facilitation) Act, 2020
- Farmers (Empowerment and Protection) Agreement on Price Assurance and Farm Services Act, 2020
- Essential Commodities (Amendment) Act, 2020

Timeline of the farmer protest

- **Sept. 17, 2020:** Ordinance is passed Lok Sabha
- **Sept. 20, 2020:** Ordinance is passed Rajya Sabha
- **Sept. 24, 2020:** Farmers in Punjab announce a three-day rail roko
- **Jan. 26, 2021:** Clash with police during tractor parade
- **Oct. 3, 2021:** Lakhimpur Kheri violence

Introduction

Active learning is a technique for reducing manual annotation effort during training phase of machine learning. The annotation is done by a human (called oracle) which helps AL systems to achieve high accuracy with few labelled instances.

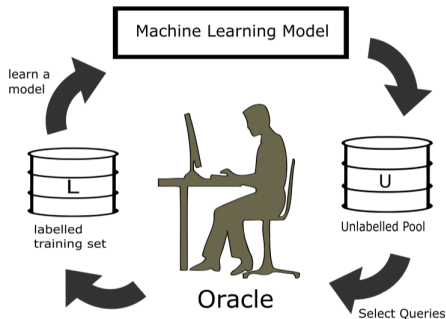


Figure: Pool-based Active Learning

- Collecting large amount of unlabeled data is easier but manually labeling them is tough.
- Active Learning reduces this effort by taking help from user and training on remaining unlabelled points.

Introduction

Active Learning Scenarios

- Membership Query Synthesis
- Stream-Based Selective Sampling
- Pool-Based Sampling

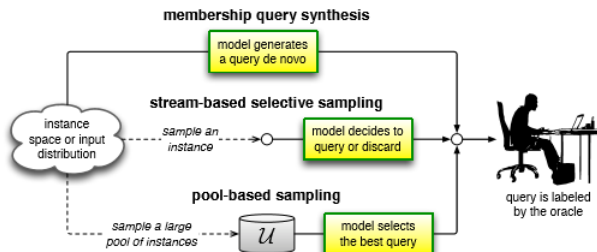


Figure: Diagram illustrating the three main active learning scenarios [10]

Motivation

Background

Voting is an important part of democracy as it allows people choose their representative. Just like citizens choose their representative, they also agree or disagree with government policies. By analyzing comments from social media sites, we can understand the sentiment of the country.

Reason for choosing Active Learning

The motivation behind the selection of Active Learning for comment classification is its high performance when problem is specific. Manual labeling of thousands of comments will take lot of time, and hence, active learning is chosen to speed up the process with accuracy.

Objectives

Objectives Specification

- i. Create a dataset by parsing comments from social media sites like YouTube.
- ii. Filter the dataset by removing unwanted characters/words.
- iii. Construct a seed set containing positive and negative comments.
- iv. Expand the seed set by randomly sampling comments from the unseen corpus.
- v. Obtain real valued embeddings for the comments.
- vi. Find the nearest neighbors (NN) of the seed set and include them in the corpus.
- vi. Expand using minority-class certainty sampling (if any).
- vii. Perform final expansion using uncertainty sampling.

i. **Enhanced LSTM for Natural Language Inference**

Chen et al. [2] proposed a state-of-the-art result on the Stanford Natural Language Inference Dataset using Long Short-Term Memory (LSTM). They employed Bi-directional LSTM (BiLSTM) as one of the building blocks. Later it is used to perform inference composition to construct the final prediction. The model has an accuracy of 88.6 %.

ii. **Semantic sentence matching with densely-connected recurrent and co-attentive information**

Kim et al. [3] propose a densely-connected co-attentive recurrent neural network to find semantic relation between sentences. To overcome the problem of ever-increasing size of feature vector due to densely connected networks, they also have propose an autoencoder after dense concatenation.

iii. **Active Learning Using Pre-clustering**

Nguyen and Smeulders [5] incorporated clustering into active learning. The algorithm first constructs a classifier on the set of the cluster representatives, and then with the help of a local noise model, it passes the classification decision to the other samples. The model allows selecting the most representative samples as well as avoids labelling samples in the same cluster. The paper focuses on discriminative models including logistic regression and Support Vector Machines (SVM) which are less sensitive to training data and hence, good for active learning.

iv. **Active Sentence Learning with AUSDS**

Ru et al. [8] propose adversarial uncertainty sampling in discrete space (AUSDS) which retrieves informative unlabeled samples more efficiently and is 10x faster when compared to typical uncertainty sampling method for active learning.


v. **Active Learning via Membership Query Synthesis for Semi-supervised Sentence Classification**

Schumann and Rehbein [9] showed that it is possible to use Membership Query Synthesis [5] for generating AL queries for natural language processing, using Variational Autoencoders for query generation, and provides competitive performance to pool-based AL strategies while substantially reducing annotation time.

vi. **Adversarial Active Learning based Heterogeneous Graph Neural Network for Fake News Detection (AA-HGNN)**

Ren et al. [7] propose a novel fake news detection framework "Adversarial Active Learning based Heterogeneous Graph Neural Network for Fake News Detection (AA-HGNN)", which employs a novel hierarchical attention mechanism to perform node representation learning in the HIN. In this paper, the authors model the news content and related entities as a News-HIN. The AA-HGNN utilizes both structural information as well as News-HIN to identify fake news.

vii. **Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus**

Xiang et al. [14] proposes a novel approach which exploits linguistic regularities in profane language via statistical topic modeling on a huge Twitter corpus, and detects offensive tweets using these automatically generated features. This approach works with various Machine Learning models such as J48 decision tree learning, SVM, 

viii. **Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection**

Watanabe et al. [13] proposes a pragmatic approach to collect hateful speech. The proposed approach uses unigram and patterns that are automatically collected from training dataset. Accuracy of 87.4% was achieved on detecting whether a tweet is offensive or not (binary classification) and 78.4% accuracy when detecting a tweet is hateful, offensive, or clean (ternary classification).

ix. **Voice for the Voiceless: Active Sampling to Detect Comments Supporting the Rohingyas**

Palakodety et al. [6] proposes a classifier which can classify comments supporting the Rohingyas. This is done by building a corpus from YouTube comments and applying multiple AL strategies based on nearest-neighbors in the comment-embedding space. The classifier provided an accuracy of 75.38% with SVM(n gram) and 77.71% with SVM(n gram + embedding).

Dataset

The dataset was constructed by parsing comments from 863 unique videos from YouTube. Total 45,376 comments were fetched out of which 40,791 comments were in pure English words. Although most of the comments uses English alphabets, it has been observed that, they are mostly Hinglish (Hindi written using English characters), and hence, for now we are taking them in consideration.

videoid	title	channelid	channelName
lrE3zoaCc	Farmers Protest: Haryana के Karnal में किसानों की बैठक से पहले बड़ी सभाओं पर रोक	SNLevX5e	NDTV India
hWxJ0YQ	Prime Time With Ravish Kumar: Can Farmer Protests Wipe Away The Stain Of Riots In Muzaffarnagar?	iMw0F81Z	NDTV
aQsQRqM	जाट वोट बैंक पर कब्जे की जंग Farmers' Protest Kisan Mahapanchayat Latest News Hindi News	i910QMde	Zee News
uLoM3Ty	Farmers Protest: Whose future will be made? Kisaan Mahapanchayat ICH	if-RFENb	ABP NEWS
/1Tu6lsID-	Farmers' Protest Is Kisan mahapanchayat turning into political stage? Debate	if-RFENb	ABP NEWS
04JAIEvZE	Farmers protest intensifies in Karnal ICH(07.09.2021)	if-RFENb	ABP NEWS
zTnA8hsY	किसानों की महापंचायत में पीएम मोदी के खिलाफ नारेबाजी Farmers' Protest Rakesh Tikait Latest News	i910QMde	Zee News
UBIfyFOYA	Farmer Protest में शामिल 9 साल के अंगद ने कहा, "में PM बना तो किसानों का राज चलेगा देश में"	NmqL72W	ABP NEWS HINDI
Op9HJAJ_	Exclusive: What will be future of Farmers' protest? Will Rakesh Tikait fight UP Election?	if-RFENb	ABP NEWS
el2RRs5Jc	#India farmers protest over new laws benefitting big firms	iBqNL5Zz	Al Jazeera English
7NjYJPwn.	Visuals of massive farmers protest in Karnal Master Stroke	if-RFENb	ABP NEWS
aWqdOxN	आज किसानों की करनाल में महापंचायत Farmers' Protest Kisan Mahapanchayat Latest News Hindi News	i910QMde	Zee News
Vp4WZ17	Haryana Farmers Protest: Officer Who Asked Cops To "Crack Heads" Of Farmers To Face Action	iMw0F81Z	NDTV
yKAgmeb	Rashtra Ki Baat: Farmers Protest में 'Allah Hu Akbar' और 'Har Har Mahadev' का क्या काम ? Manak Gupta	2AYqHcLV	News24
qFno8rjO6sga	farmer mahapanchayat by protesters in UP ahead of polls, Rakesh Tikait's 'graveyard' warniit2yB_WzL		Hindustan Times

Figure: Snapshot of Dataset containing search results

Methodology

Comment	Likes	Video ID
किसान मजदूर एकता जिन्दाबाद	0	0lrE3zoaCql
Mam actually ek kisan ki death ho gyi hai usko police walo ne itna mara	0	0lrE3zoaCql
Ko bi	0	0lrE3zoaCql
किसान भाईयों यदि पुलिस वाले लाठी चलाने की कोशिश करें तो उनकी लाठी छीन लेना और जमा कर लेना	0	0lrE3zoaCql
Very bad BJP cm jjp	0	0lrE3zoaCql
पानों को बहुत रोना पड़ेगा राकेश टिकैत भारत में जिहाद और शरिया कानून लागू करना चाहते हैं खालिस्तान और पाकिस्तान से पैसे लेकर अल्ला ह	0	0lrE3zoaCql
Kisan kalank Tikait ko bhejo jail.	0	0lrE3zoaCql
ना काल में इतनी भीड़ क्यों इकट्ठा कर रहे हो नेताओं। जो कहना है वीडियो बनाकर डाल दो। लोग देख लेंगे। क्यों लोगो की जान जोखिम में डाल र	1	0lrE3zoaCql
Kisan zindabad	0	0lrE3zoaCql
Rasoia Kuta murdabad, randwo ki najyaj olad	0	0lrE3zoaCql
Khichari khud chor mandli ka sargana gujrati lutera ke liye 🙄	0	0lrE3zoaCql
T	0	0lrE3zoaCql
-j	0	0lrE3zoaCql
Kisan Ekta zindawad	0	0lrE3zoaCql
Jai Kisan	0	0lrE3zoaCql

Figure: Snapshot of Dataset containing user comments

Video and comment statistics

On analyzing unigram [kisan] and bigram [kisan nahi], clouds of 788 and 45 unique words were formed. Table 1 shows the most frequent words of these two clouds.

kisan	frequency	kisan nahi	frequency
ekta	844	dalal	5
jai	228	aatankwadi	3
majdoor	183	khalistani	3

Table: Most Frequently Occurring Word in Cloud

Seed words

To start the active learning process, we will require some manual seed words. Table 2 shows the manually selected seed words.

Positive

support
love
power
zindabad
care
jay jawan jay kisan
proud

Negative

evil
shame
dacoit
illegal
riot
khalistani
selfish

Table: Manually Selected Seed Word

User level analysis

After analyzing various comments and YouTube videos, we have observed that there are two groups of channels that either make videos in favor of farmers or against of farmers. Table 3 shows sentiment of media houses towards farmers.

For	Against
NDTV	Zee News
BBC News	Republic TV
Al Jazeera	Times Now

Table: Sentiment of Media Houses Towards Farmers

Conclusion

Conclusion

In this report we provided active learning based approach to classify comments supporting the farmers protest. We also discussed the related researches in the field of Active learning. As we are working with our own dataset, it gives much more flexibility and provides latest results. The statistical analysis showed us the sentiments of people towards farmers.

Scope for further research

The report is limited to Data pre-processing. As for future work, we will be completing the remaining objectives.

References I



Dana Angluin.

Queries and concept learning.

Machine learning, 2(4):319–342, 1988.



Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen.

Enhanced lstm for natural language inference.

arXiv preprint arXiv:1609.06038, 2016.



Seonhoon Kim, Inho Kang, and Nojun Kwak.

Semantic sentence matching with densely-connected recurrent and co-attentive information.

In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6586–6593, 2019.

References II



David D Lewis and William A Gale.

A sequential algorithm for training text classifiers.

In *SIGIR'94*, pages 3–12. Springer, 1994.



Hieu T Nguyen and Arnold Smeulders.

Active learning using pre-clustering.

In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004.



Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell.

Voice for the voiceless: Active sampling to detect comments supporting the rohingyas.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 454–462, 2020.

References III



Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang.

Adversarial active learning based heterogeneous graph neural network for fake news detection.

In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 452–461. IEEE, 2020.



Dongyu Ru, Jiangtao Feng, Lin Qiu, Hao Zhou, Mingxuan Wang, Weinan Zhang, Yong Yu, and Lei Li.

Active sentence learning by adversarial uncertainty sampling in discrete space.

arXiv preprint arXiv:2004.08046, 2020.

References IV



Raphael Schumann and Ines Rehbein.

Active learning via membership query synthesis for semi-supervised sentence classification.

In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 472–481, 2019.



Burr Settles.

Active learning literature survey.
2009.



Burr Settles, Mark Craven, and Lewis Friedland.

Active learning with real annotation costs.

In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:, 2008.



Amar Shankar.

Indian agriculture farm acts: 2020.

International Journal of Modern Agriculture, 10(2):2907–2914, 2021.



Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki.

Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection.


IEEE access, 6:13825–13835, 2018.



Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose.

Detecting offensive tweets via topical feature discovery over a large scale twitter corpus.

In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984, 2012.

-  [Xiaojin Jerry Zhu.](#)
Semi-supervised learning literature survey.
2005.

Thank you!!