Classification of Comments Supporting the Farmer Protest

Ajay Biswas

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Classification of Comments Supporting the Farmer Protest

Thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Technology

in

Computer Science and Engineering
(*Specialization*: *Information Security*)

by

Ajay Biswas

(Roll Number: 220CS2184)

based on research carried out

under the supervision of

Prof. Tapas Kumar Mishra

May, 2022

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Department of Computer Science and Engineering
National Institute of Technology Rourkela

May 06, 2022

Certificate of Examination

Roll Number: *220CS2184*
Name: *Ajay Biswas*
Title of Dissertation: *Classification of Comments Supporting the Farmer Protest*

We the below signed, after checking the thesis mentioned above and the official record book (s) of the student, hereby state our approval of the thesis submitted in partial fulfillment of the requirements of the degree of *Master of Technology* in *Computer Science and Engineering* at *National Institute of Technology Rourkela*. We are satisfied with the volume, quality, correctness, and originality of the work.

_____
Tapas Kumar Mishra

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Prof. Tapas Kumar Mishra
Professor

May 06, 2022

Supervisor's Certificate

This is to certify that the work presented in the thesis entitled *Classification of Comments Supporting the Farmer Protest* submitted by *Ajay Biswas*, Roll Number 220CS2184, is a record of original research carried out by him under my supervision and guidance in partial fulfillment of the requirements of the degree of *Master of Technology* in *Computer Science and Engineering*. Neither this thesis nor any part of it has been submitted earlier for any degree or diploma to any institute or university in India or abroad.

_____
Tapas Kumar Mishra

Dedication

I dedicate this thesis to my parents. Without their love and support, the completion of this work would not have been possible.

Signature

Declaration of Originality

I, *Ajay Biswas*, Roll Number *220CS2184* hereby declare that this thesis entitled *Classification of Comments Supporting the Farmer Protest* presents my original work carried out as a postgraduate student of NIT Rourkela and, to the best of my knowledge, contains no material previously published or written by another person, nor any material presented by me for the award of any degree or diploma of NIT Rourkela or any other institution. Any contribution made to this research by others, with whom I have worked at NIT Rourkela or elsewhere, is explicitly acknowledged in the dissertation. Works of other authors cited in this dissertation have been duly acknowledged under the sections "Reference" or "Bibliography". I have also submitted my original research records to the scrutiny committee for evaluation of my dissertation.

I am fully aware that in case of any non-compliance detected in future, the Senate of NIT Rourkela may withdraw the degree awarded to me on the basis of the present dissertation.

May 06, 2022
NIT Rourkela

Ajay Biswas

Acknowledgment

This section shows the essence of the student's journey during the course of the research work and the role of other individuals in shaping his/her academic life. Acknowledgments are non-consequential in that a student is not evaluated on them. An acknowledgment has typically three sections —

- *Reflection*: Narration of the student's journey through his/her research career.
- *Thanking*: Expression of gratitude to those who have helped in the student's journey.
- *Announcement*: Accepting responsibility for the work and/or dedication of the dissertation to someone. This in fact is a repeat of declaration and dedication pages.

May 06, 2022 *Ajay Biswas*
NIT Rourkela Roll Number: 220CS2184

Abstract

The introduction of farm bills 2020 was seen as a major agricultural reform, however, started a year-long protest which finally ended with the repeal of the bills. In this research we tried to classify YouTube comments which are in the support of the farm bills or are against of it. We used Active Learning and Weak Supervision to classify comments which works well with real world situation. We used FastText for generating Word Embeddings and did analysis using four popular classifiers. TO BE COMPLETED.

Keywords*: *active learning*; *farmer protest*; *weak classification*; *fasttext*; *uncertainty sampling

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The Farm Bills, or the Indian agriculture acts of 2020, are three acts initiated by Parliament of India during September 2020. The three farm acts are as follows: ”Farmers’ Produce Trade and Commerce (Promotion and Facilitation) Act, 2020; Farmers (Empowerment and Protection) Agreement on Price Assurance and Farm Services Act, 2020; and Essential Commodities (Amendment) Act, 2020”. Although the bills were supposed to be beneficial for the farmers, they led to a mass protest, which gained momentum in September 2020 [1]. The Protest is not limited to grounds, but also spread across various social media websites like YouTube, Reddit, Facebook, Instagram and Twitter. Our main goal is not to argue who is right or wrong but to identify the comments that are in favor and against of the Indian Farmers’ protest.

Our proposed approach is inspired from the work done in [2]. This paper proposed an Active Learning (AL) based classifier that can classify comments supporting the Rohingyas. Further discussions related to this paper will be dealt in the literature survey section. 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. For problems having large collection of unlabeled data, pool-based sampling is used [3] as shown in figure 1.1. AL is highly useful in classifying comments which involves a person’s opinion, belief or political interest, as it’s very tough to label large number of comments accurately and effortlessly by a human. Also, there are numerous challenges to be dealt with before any classification could take place. Some of the challenges are (i) Dealing with multiple languages having different levels of grammatical accuracy, (ii) Un-structured data, (iii) ambiguous sentences, (iv) Unrelated comments, etc.

1.2 Applications of Active Learning

Active learning is becoming a burning topic in the field of machine learning due to its high performance and wide range of uses. Some of the areas where active learning can be useful are as follows:

- *Speech Recognition.* One by one labeling speech utterances can be very challenging [4] as well as time consuming. As speech may have several languages or multiple dialects, trained linguists are needed for this purpose.
- *Information Extraction.* To extract high quality information, hours of manual labor is required. For highly specialized job, professional are required like Genomic information retrieval requires Phd-level candidates [5].

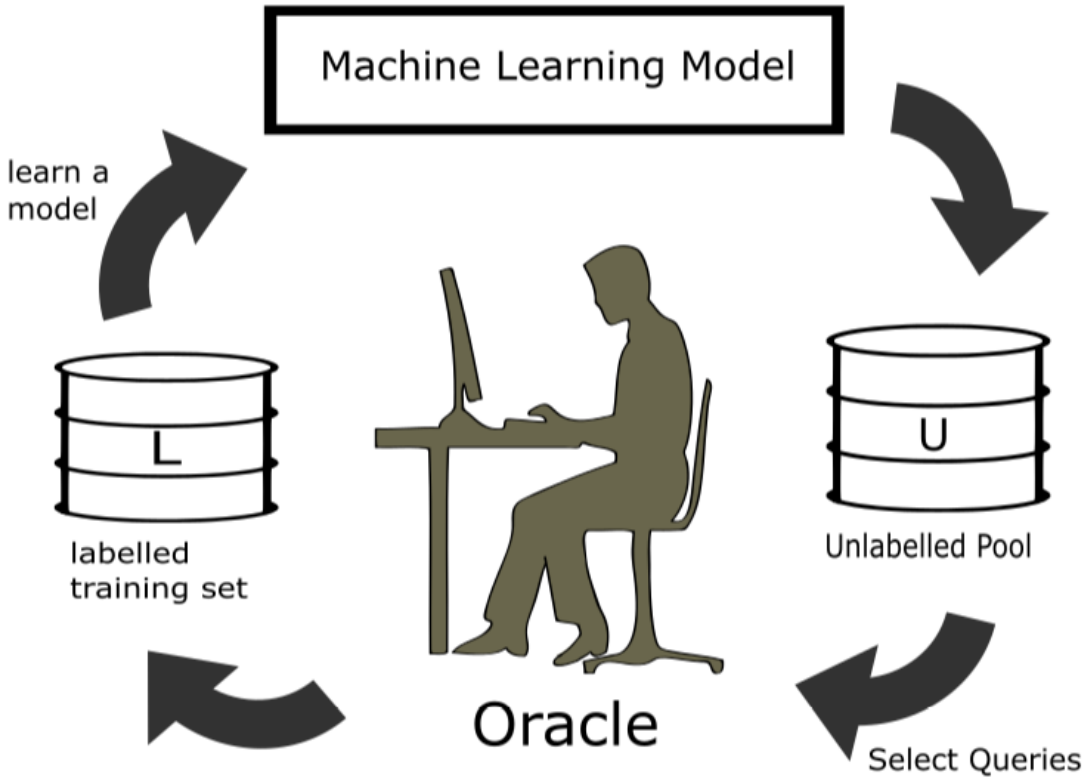


Figure 1.1: Pool-based Active Learning

1.3 Word Embedding Techniques

Humans can read sentences and understand their meaning, but computers don't work this way. The text data has to be transformed into numerical data which they will use for

classification. These numerical data are known as Word Embeddings. This is an important step before classification and can give fascinating results if a good technique is used. We will be extensively using the following two word embedding techniques in our project to generate word embeddings.

1.3.1 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a technique which quantifies words in a set of documents. It is used to measure importance of words in data. The Term Frequency (TF) is the number of times a word appeared in a document divided by total number of words in that particular document. Inverse Document Frequency (IDF) is the document frequency of the word among all documents. The formula of TF-IDF is shown in equation 1.1.

$$\begin{aligned} TF(t,d) &= \frac{f_d(t)}{|d|} \\ IDF(t,D) &= \ln \frac{|D|}{|\{d \epsilon D : t \varepsilon d\}|} \\ TF(t,d) - IDF(t,D) &= TF(t,d) \times IDF(t,D) \end{aligned} \tag{1.1}$$

Where $f_d(t)$ is the number of times term t appeared in document d, $|d|$ is the number of words in document d, and $|D|$ is the total number of documents.

1.3.2 FastText

FastText [6] is an open-source library from Facebook for word embeddings and text classification. It uses hierarchical classifiers to build model which is faster compared to the models built through deep neural networks which uses linear classifiers. FastText provides both supervised and unsupervised learning and supports both Continuous Bag of Words (CBOW) and Skip-gram models. For unsupervised learning, where labels are not available, FastText uses N-Gram Technique to split words into n-gram components which captures meaning of suffixes and prefixes. FastText can also provides embeddings of Out of Vocabulary words which gives it upper edge than other competitors.

1.4 Motivation

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. The motivation behind the selection of Active Learning for comment classification is its high performance when problem is specific.

1.5 Organization of the Thesis

The organization of this thesis is as follows: Chapter 1 provides brief introduction briefly describes about the ongoing farmer protest and how active learning can be used to classify comments supporting the protest. It also outlines the motivation and objective of this work; Chapter 2 provides a brief summary on the related works are summarized in this chapter; Chapter 3 describes our proposed work. It describes how we built the dataset, how we applied active learning and weak supervision to make labelling process faster as well as compared the findings with one similar work done in the field of Active Learning. Finally, Chapter 4 presents the conclusion our this work, the limitations and the areas that can be improve in future work.

Chapter 2

Literature Survey

2.1 Overview

This chapter contains a brief survey of research work done in the field of Sentence Classification and Active Learning.

2.2 Related Works

Previously various research were conducted in the field of Active Learning and Sentence Classification. We are focusing on those researches which are related to our work.

i. Chen et al. [7] 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.

ii. Kim et al. [8] 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. Nguyen and Smeulders [9] 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. Ru et al. [10] 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. Schumann and Rehbein [11] 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. Ren et al. [12] 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. Xiang et al. [13] 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, logistic regression (LR) and random forest (RF).

viii. Watanabe et al. [14] 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. Palakodety et al. [2] 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.

Chapter 3

Classification of Comments Supporting the Farmer Protest

3.1 Overview

This chapter provides the proposed work done so far in classification of comments supporting the Indian farmers' protest.

3.2 Dataset

To construct the dataset, first 4,12,445 comments were parsed from the comment section of YouTube using the YouTube API which came from 1077 unique videos. A Corpus of 16,842 comments was made by using a language filter which keeps texts of purely English characters and a keywords based filter which chooses comments which have those keywords or their different spelling variations. This removes most of the ambiguous and spam comments from the dataset. Some of the keywords are given in the table 3.1. These comments maybe written in either pure English or in some other language written in Latin script like Hinglish which is Hindi written using English alphabets.After that basic text pre-processing is done like contraction expansion, punctuation, url, username, and emoji removal, converting to lowercase and lemmatization. We did not removed stop words as fastText can handle them and some of them signifies stance like ”against” and ”not”. as To make a labelled dataset for Machine Learning, 4478 comments were hand labelled, out of which 1846 labelled 0, 1408 labelled 1, and 1224 labelled 2, which signifies comments Against the Farm Bills, In Support of the Farm Bills and Neutral, respectively. Out of these three labels, we considered comments of either label 0 or 1, i.e. either against or in support of the bill. To make the labelling process more user friendly and to make comments of uniform length, we trimmed each comments up to 300 characters.

Table 3.1: Keyword Based Filtering List

Keywords
aatankwadi, against, apmc, bill bjp, bsp, congress, choukidar dacoit, dakaat, dakat, election farm, farmer, godi, haryana, india, kejriwal, khalistan, kisaan lakhimpur, majdoor, mandi, media modi, modiji, msp, pakistan price, punjab, rakesh, rally ravish, reject, repeal, rss rubbish, sapot, sarkar,support taliban, terrorist, tikait tractor, victory, yogi, yogiji zee, zindabad

3.3 Dataset Labelling

Dataset labelling is one of the most difficult and time consuming task in the entire Supervised Machine Learning process. Active Learning solves this problem by automating labelling process along with the user. Since we are trying to find best results in classifying the comments, we have to label entire dataset to compare the accuracy.

Labelling comments subjected to political debates are challenging as the person has to possess knowledge of current politics, laws as well as grammar. Lack of these may result in biasness in dataset. To minimize this effect, the labelling was done and cross checked by the authors of this paper. Table 3.2 shows the labels assigned to the comments.

Table 3.2: Comments Along With Their Assigned Labels

No.	Comments	Labels
1.	Farmers protest is revolutionary and historical.	0
2.	Modhi don't like farmers. He want them be silent.	0
3.	Modiji affraid of UP Election	0
4.	The actual farmers are working in the field....and contributing to the nation	1
5.	I m support of bill	1
6.	Its not really anti farmers but anti middleman	1

Figure 3.1: Wordcloud of the Dataset

3.4 Dataset Analysis

Dataset analysis is an important step before any kind of classification. Although the dataset is formed by comments from the people having a serious bone of contention, one thing is common, either they are in support or against of the bill. Figure 3.1 shows the wordcloud showing the most frequent words in the dataset. The stance of a comment could be identified by analyzing the kind of words present in it. Figure 3.2 shows the class-wise frequency of emotional words present in the dataset. FIgure 3.3 and 3.4 shows the frequency of frequent bigrams and trigrams respectively.

Apart from emotional words, we can study the dataset by plotting the points (rows of dataset) with respect to the presence of a keyword/phrase in the comments. Figure 3.5 shows the plot of dataset comments based on TF-IDF scores of two phrases. Since the words

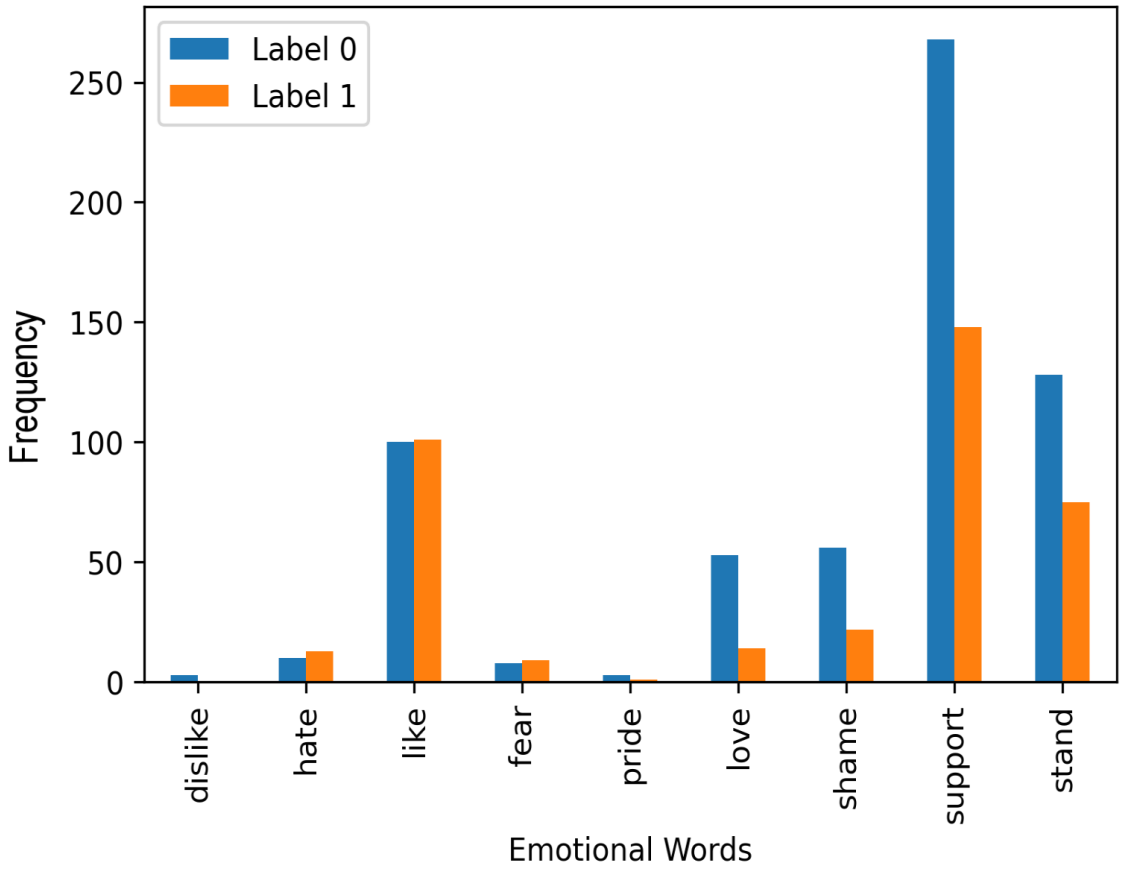


Figure 3.2: Frequency of Emotional Words in the Dataset

may or may not be present in each comments, they are not reliable source of classification criteria. They will be used in Weak Supervision which will be later discussed in the upcoming sections.

3.5 FastText Unsupervised Model

The corpus which we made after text pre-processing, was fed to fastText for unsupervised learning to generate two unsupervised models of Word N-Grams (N=2 and N=3). The model's parameters were 300 dimensions, 0.01 learning rate and 50 epochs. All the analysis shown in the upcoming sections are done using the N=2 model.

3.6 Metrics and Classifiers

3.6.1 Performance Metrics

We took four metrics to evaluate performance of our technique; Accuracy, Precision, F1 score and Recall, whose formulas are given in equation 3.1, 3.2, 3.3, and 3.4 respectively.

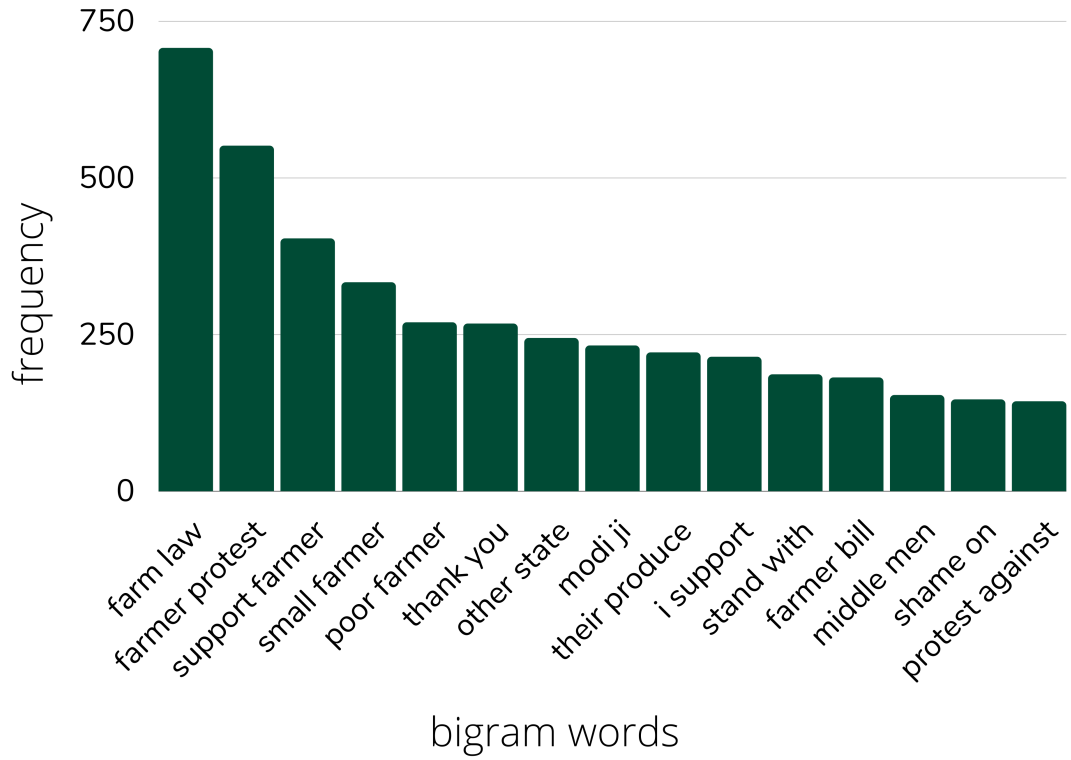


Figure 3.3: Frequency of Frequent Bigrams in the Dataset

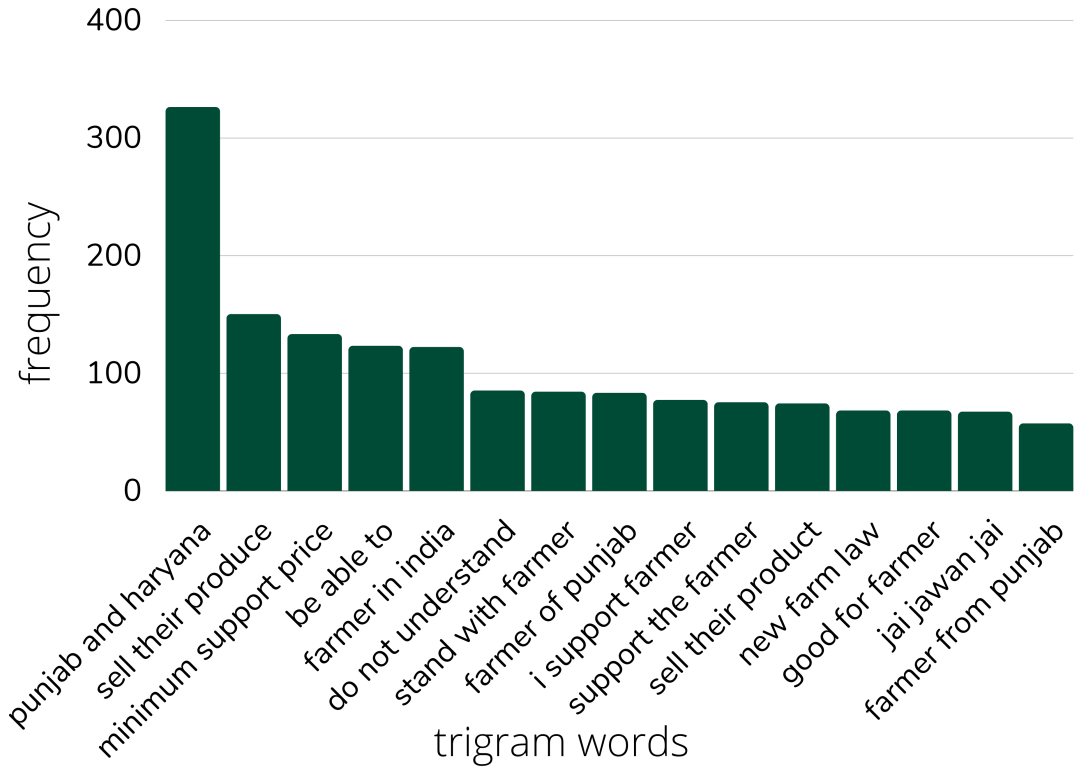
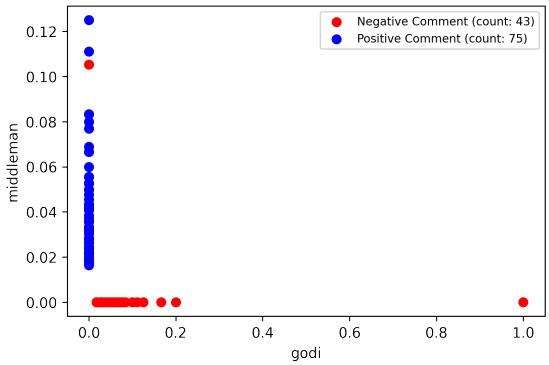
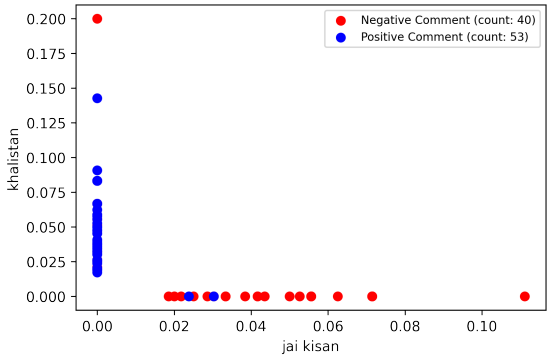


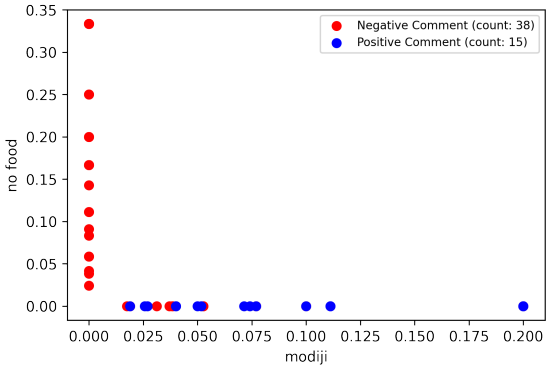
Figure 3.4: Frequency of Frequent Trigrams in the Dataset



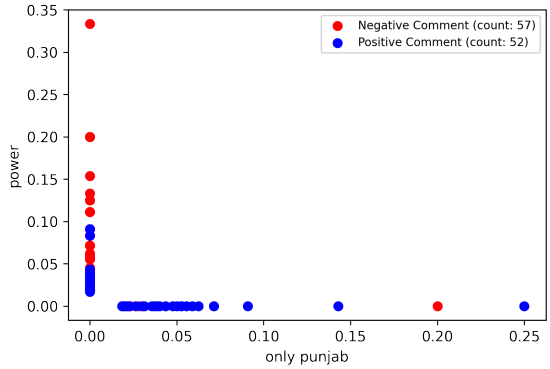
(a) Godi vs. Middleman



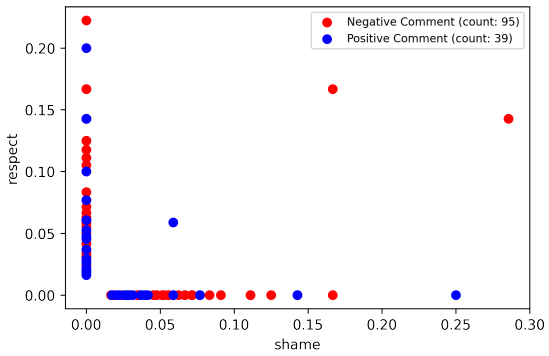
(b) Jai Kisan vs. Khalistan



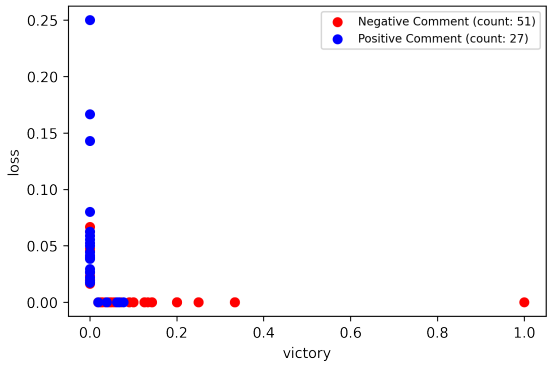
(c) Modiji vs. No Food



(d) Only Punjab vs. Power



(e) Shame vs. Respect



(f) Victory vs. Loss

Figure 3.5: Weak Classification Using TF-IDF scores of Two Phrases

These are calculated with the help of confusion matrix as shown in figure 3.6.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.3}$$

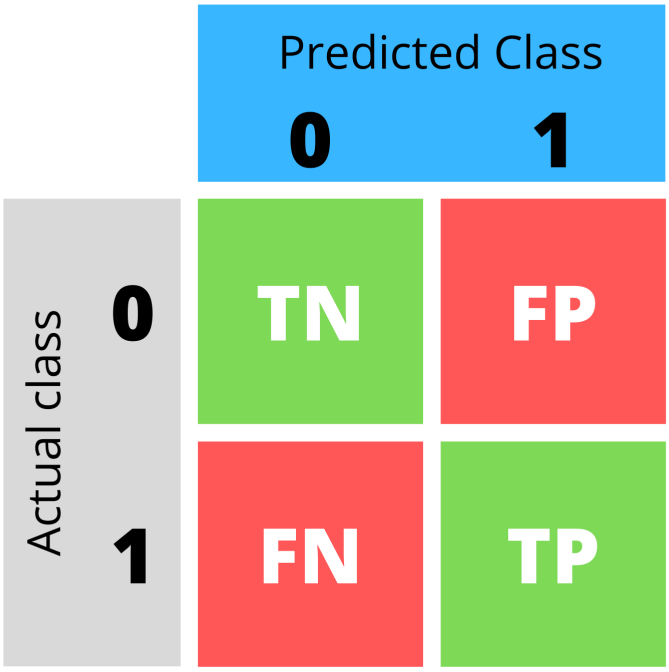


Figure 3.6: Confusion Matrix

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

Here TP, TN, FP and FN means True Positives, True Negatives, False Positives and False Negatives respectively.

3.6.2 Hyperparameter Tuning and Verification

Validation curves are used to tune single hyperparameter of a classifier. It has two curves, one accuracy curve and other k-fold cross-validation curve. The plotting is performed by varying the value of hyperparameter and plotting the accuracies and crossvalidation accuracies. A sweet spot for choosing optimal value of hyperparameter will the point where both curves have high accuracy as well as the gap between them is low. A low accuracy score denotes underfitting and high gap denotes overfitting.

3.6.3 Receiver Operating Characteristic (ROC) and Detection Error Tradeoff (DET) Curves

ROC curve is the plot of True Positive Rate vs. False Positive Rate at different classification thresholds. The Area Under Curve (AUC) is the region below the curve which measures how well the predictions are ranked.

DET curve is a plot of False Negative Rate vs. False Positive Rate. DET curve allows easier analysis of classifiers because of the linear scale. The user can directly check at which False Positive Rate, the False Negative Rate is low and vice-versa.

3.7 Execution and Results

For training and testing, we took a split of 80-20 percent respectively. Out of 2603 training samples, 1477 are negatives (have label 0), i.e. comments which are against bill, and 1126 are positives (have label 1). Then, out of 651 testing samples, 369 are negatives, and 282 are positives. We ran four classifiers which are Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) after adjusting some of their hyper-parameters by plotting Validation Curve which is shown in figure 3.10. After that we plotted ROC and DET curves with those hyper-parameters as shown in figure 3.12.

As we know, Active Learning reduces effort of manual annotation, we decided to perform two Active Learning techniques; Random Sampling and Uncertainty Sampling. First, we randomly split training data into seed set and expansion set with a split of 1-99%. We are trying to start from a tiny seed set and trying to expand it. The seed set contains 15 negatives and 11 positive comments, and expansion set contains 2577 comments.

3.7.1 Random Sampling

In this sampling technique, we took batches of 20 comments from the expanded set and pick comments one by one, found out their 40 Nearest Neighbor (NN) vector using the FastText model and calculated cosine similarity with that of seed set. Whichever comment had the highest similarity score, we will classify the comment to that class and expand the seed set. Finally we will pick randomly r number of comments and correct their assigned labels if any. The diagram of random sampling is shown in figure 3.7. The results of random sampling with $r = 4$ is shown in table 3.3 along with results of uncertainty sampling which will be discussed in the next section.

3.7.2 Uncertainty Sampling

In this sampling technique, just like random sampling, we took comments in batches of 20, but instead of finding their nearest neighbors, we first transformed both seed set and expansion set using the unsupervised model, and ran predictions on transformed testing set using seed set as training data. We sorted them based on their predicted probability score in non decreasing order using the smallest-margin [15] technique as shown in equation 3.5. Here, the most uncertain comments are picked up and are labelled by the user. Rest of the comments are thrown away because they could be classified easily and won't help

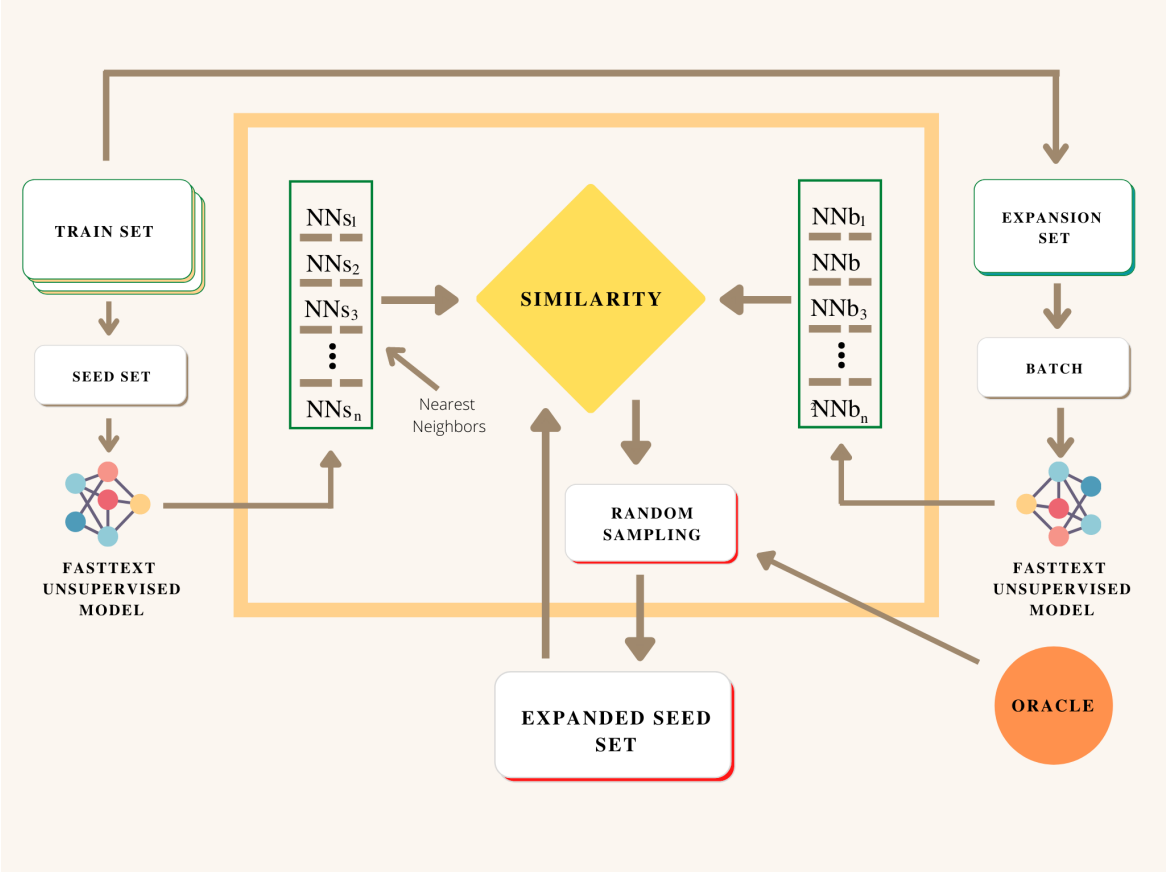


Figure 3.7: Batch-wise Nearest Neighbor Based Random Sampling

in classifying uncertain comments. The testing was performed by taking top five most uncertain comments from each batch and appending them to the seed set. The batch size and number of uncertain comments seems arbitrary, however, they were chosen after testing with different values of batch size and number of top uncertain comments, by taking accuracy, speed of execution and underfitting and overfitting into consideration. Figure 3.8 shows the batch-wise uncertainty sampling technique we used to expand the seed set. Figure 3.9 shows the comparison of confusion matrices of no active learning, random sampling and uncertainty sampling. Figure 3.11 shows the validation curve of the four different classifiers. Figure 3.12 shows the comparison of ROC and DET curve of training data with the expanded set. Figure 3.13 and 3.14 shows the ROC and DET curve of four different expanded seed sets when classified using Logistic Regression, SVM, KNN and MLP respectively. Figure 3.15 shows the accuracy, precision (weighted), recall (weighted), and f1 score (weighted) of classifying the test set at each step of expansion. Here batch size was 20 and top 5 uncertain comments were picked and labelled.

$$\phi_M(x) = P_\theta(y_0^*|x) - P_\theta(y_1^*|x) \tag{3.5}$$

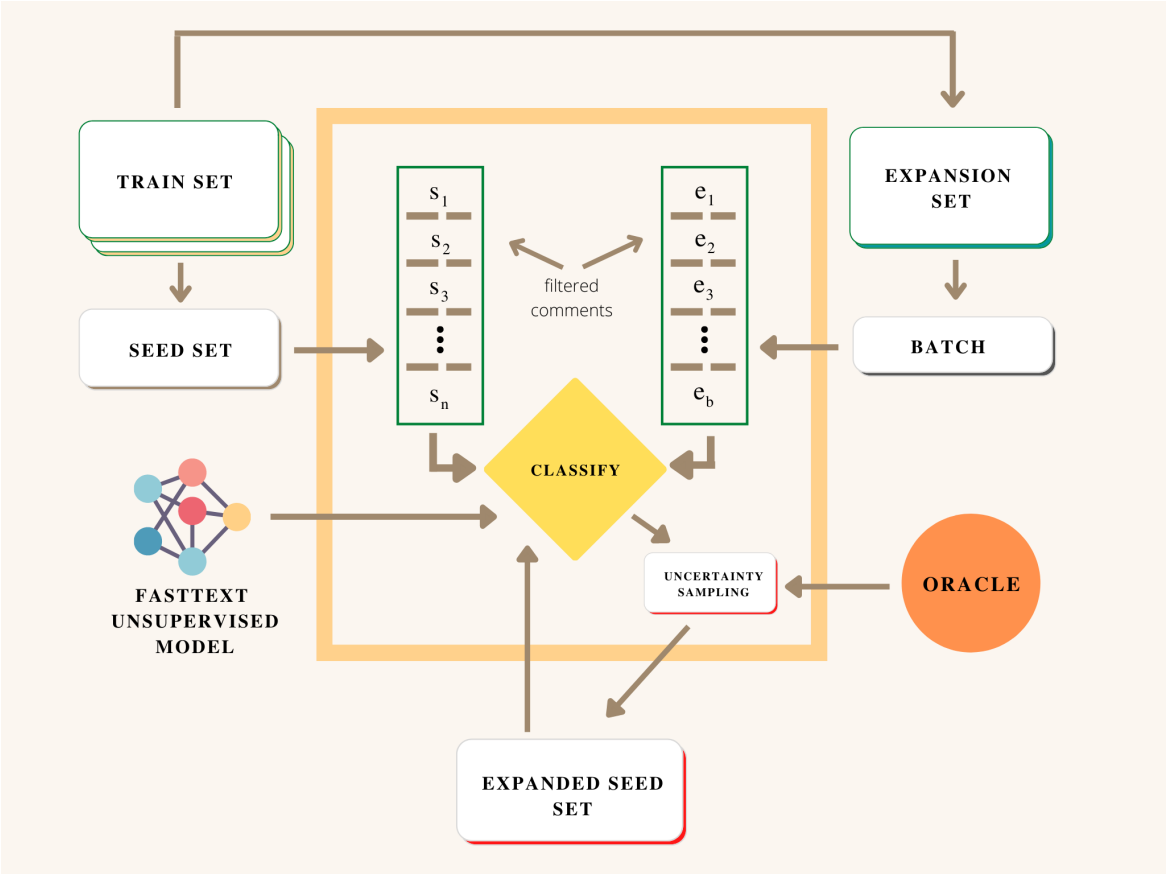


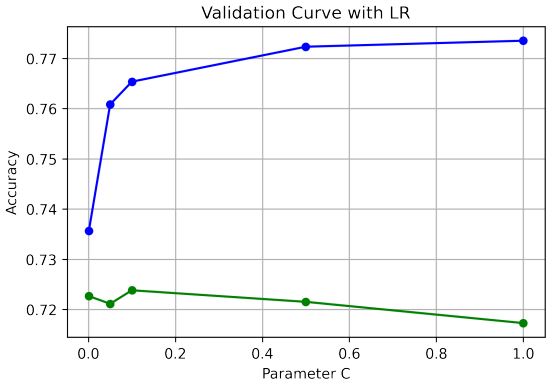
Figure 3.8: Batch-wise Word Embedding Based Uncertainty Sampling

3.7.3 Comparison With Existing Work

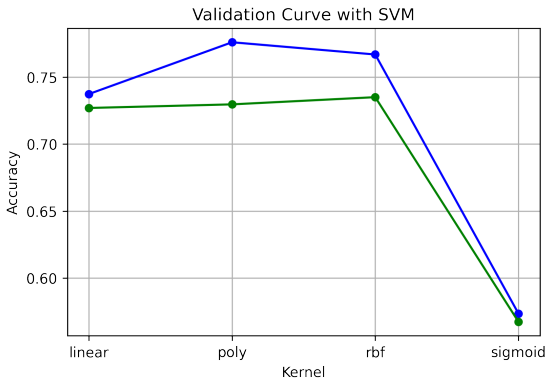
Table 3.4 summarizes the results of Palakodety et al.'s [2] work. They first begin with small seed set and then expanded using Random Sampling, Certainty Sampling and Uncertainty Sampling. Then they used SVM (n gram + embeddings) to further their results. Noting down their results, we further checked with three more classifiers with in-depth analysis with different sizes of expanded set which they lagged.

		Supervised Learning			Random Sampling			Uncertainty Sampling	
		Predicted			Predicted			Predicted	
		0	1		0	1		0	1
LR	Actual 0	293	77	0	283	87	0	301	69
	1	85	196	1	102	179	1	65	216
SVM	Actual 0	312	58	0	293	77	0	326	44
	1	79	202	1	114	167	1	69	212
KNN	Actual 0	251	119	0	236	134	0	239	131
	1	69	212	1	80	201	1	65	216
MLP	Actual 0	296	74	0	276	94	0	313	57
	1	76	205	1	102	179	1	59	222

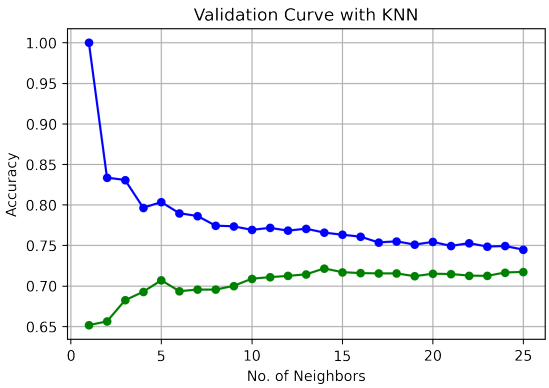
Figure 3.9: Comparison of Confusion Matrices for Three Different Techniques



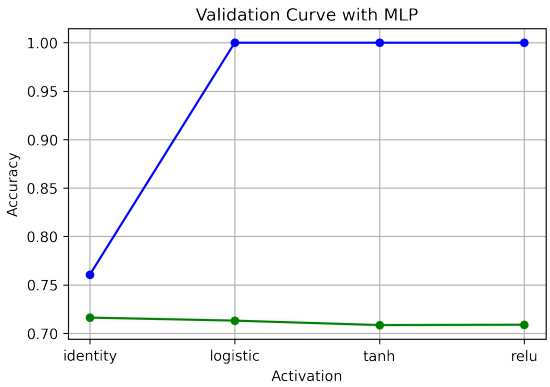
(a) Logistic Regression



(b) SVM

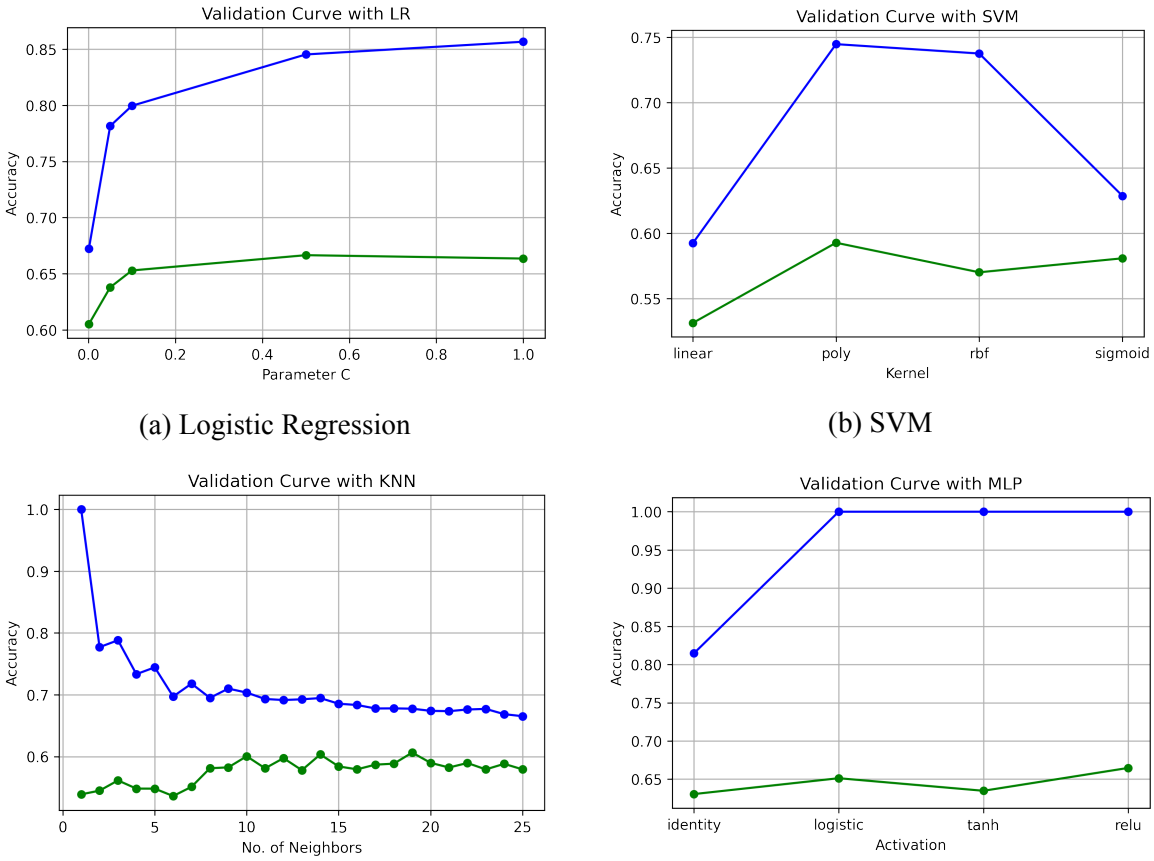


(c) KNN



(d) MLP

Figure 3.10: Validation Curves of Four Classifiers on Training Data



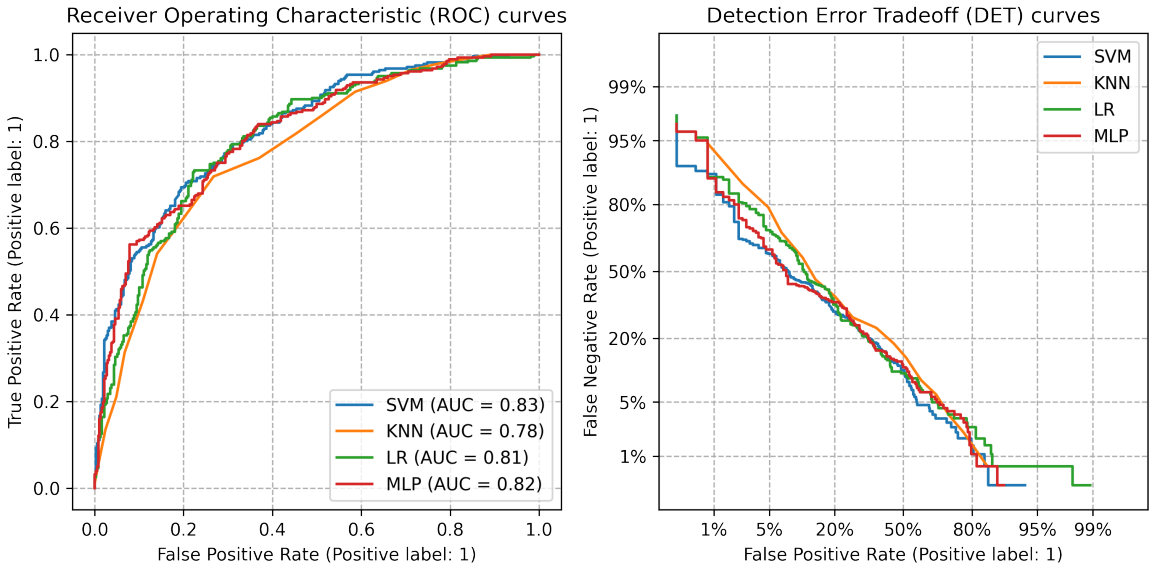
(a) Logistic Regression

(b) SVM

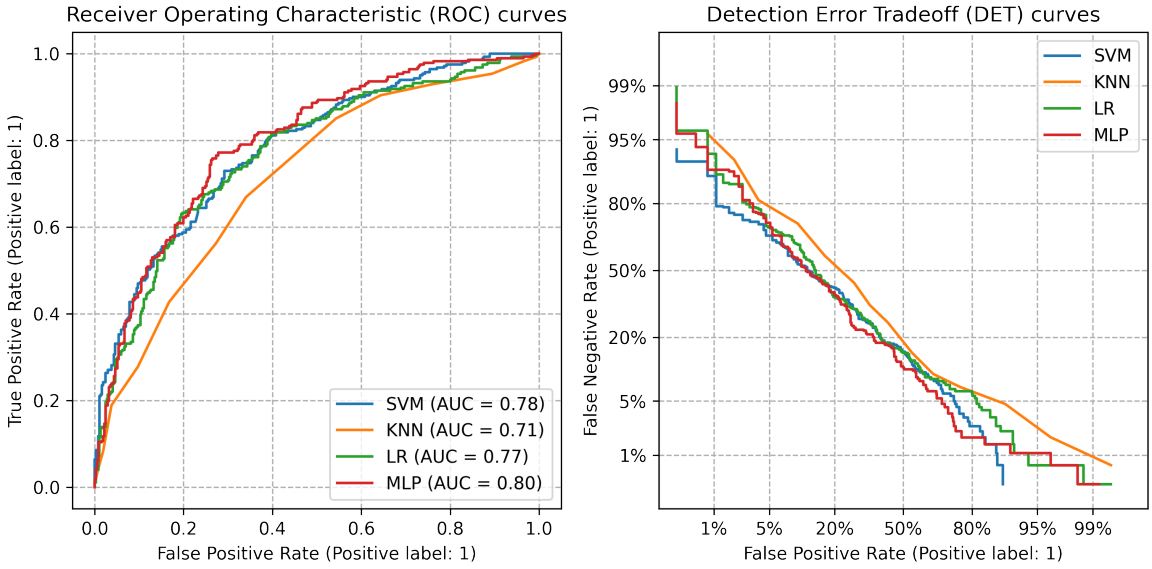
(c) KNN

(d) MLP

Figure 3.11: Validation Curves of Four Classifiers on Expanded Set

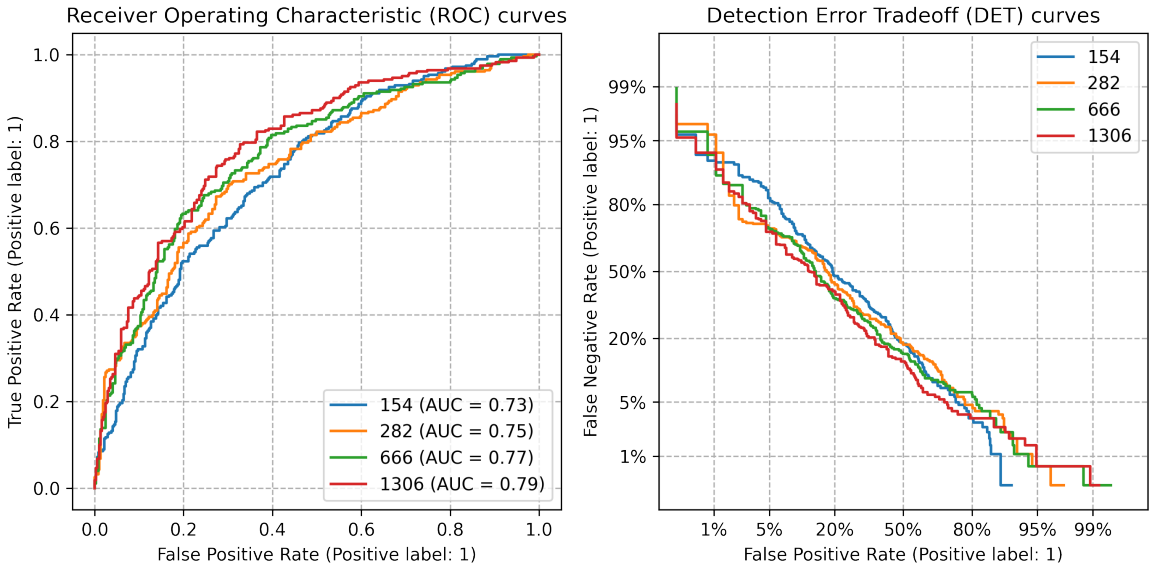


(a) Training Data

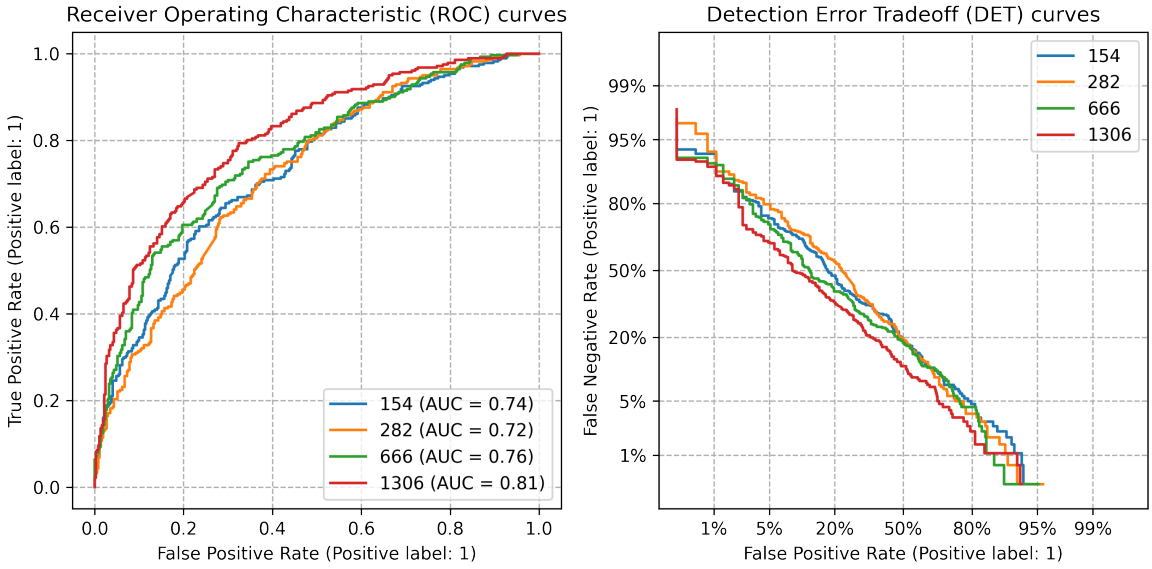


(b) Uncertainty Sampled Data

Figure 3.12: Comparative ROC and DET curves of Four Classifiers on Training and Uncertainty Sampled Data)

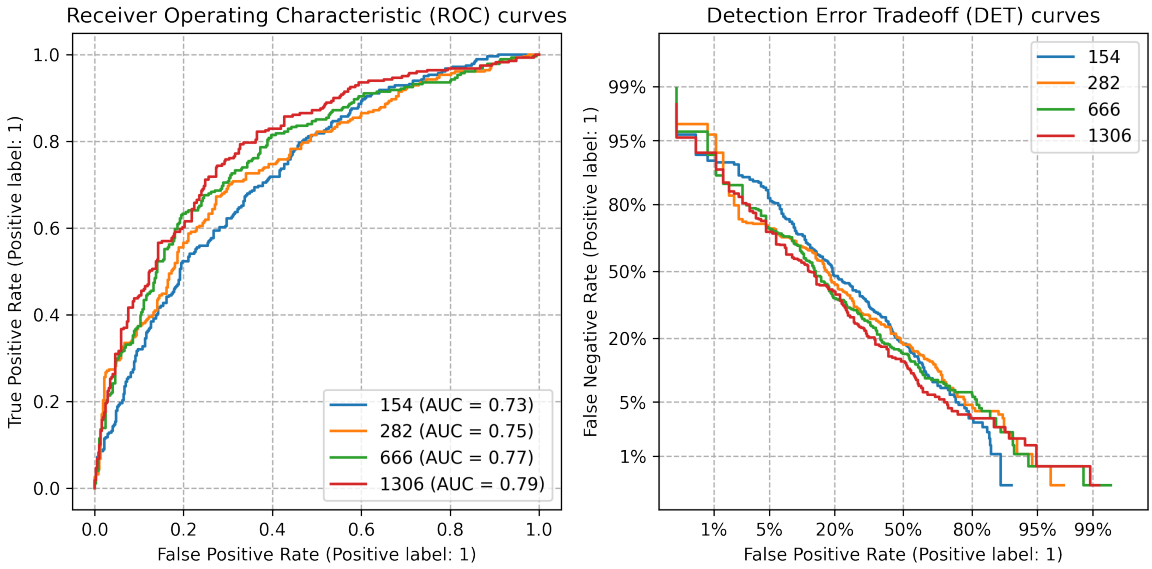


(a) Logistic Regression

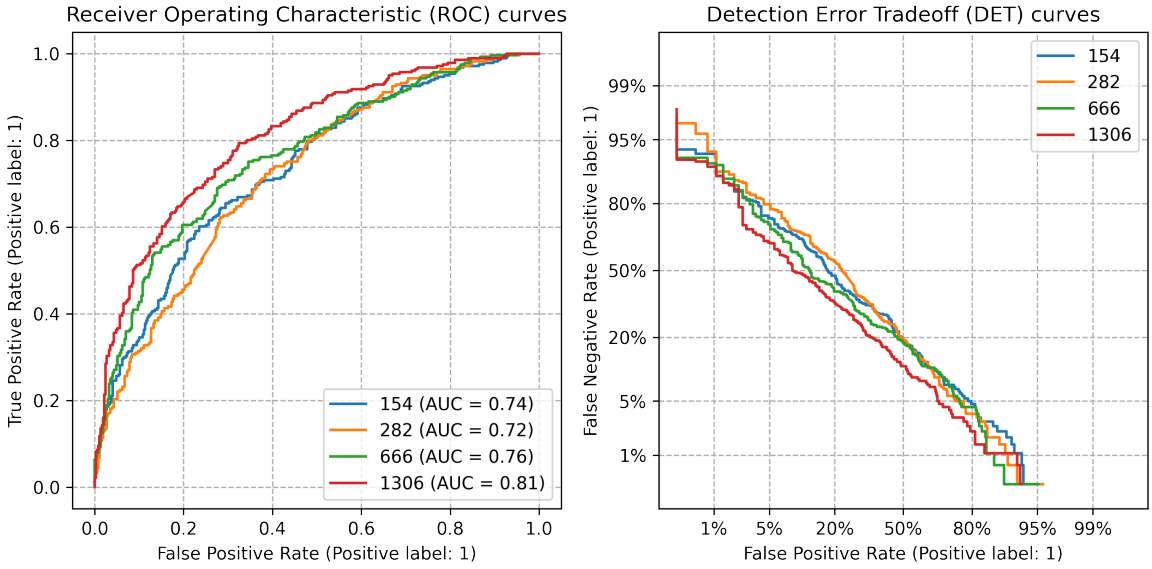


(b) SVM

Figure 3.13: Comparative ROC and DET curves of Logistic Regression and SVM on Different Sizes of Expanded Set

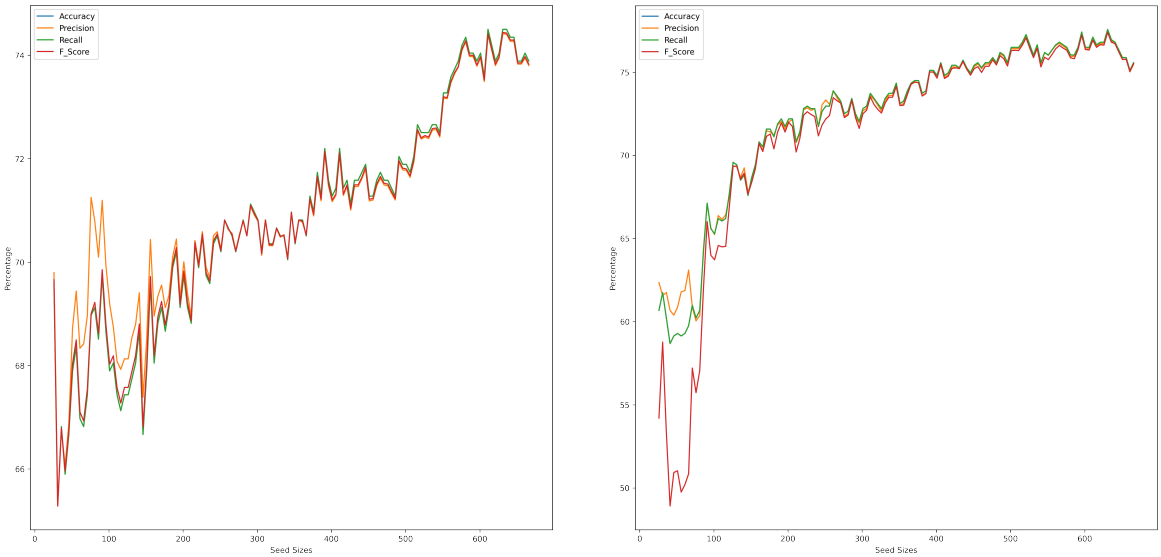


(a) KNN



(b) MLP

Figure 3.14: Comparative ROC and DET curves of KNN and MLP on Different Sizes of Expanded Set



(a) Logistic Regression

(b) SVM

(c) KNN

(d) MLP

Figure 3.15: Classification Results During Seed Set Expansion

Table 3.3: Classification Results of Batch-wise Word Embeddings Based Uncertainty Sampling

Algorithm	**WordNGrams**	**Accuracy %**	**Precision %**	**F1 score %**	**Recall %**
LR	N=2	75.12	71.80	69.75	70.76
SVM	N=2	78.96	77.69	71.88	74.68
KNN	N=2	71.12	64.04	75.44	69.29
MLP	N=2	76.96	73.48	72.95	73.21
LR	N=3	75.88	72.97	70.11	71.51
SVM	N=3	78.80	76.38	73.67	75.00
KNN	N=3	71.12	63.80	76.51	69.58
MLP	N=3	77.11	73.57	73.31	73.44
LR + Random + NN	N=2	70.97	67.30	63.70	65.45
SVM + Random + NN	N=2	70.66	68.44	59.43	63.62
KNN + Random + NN	N=2	67.13	60.00	71.53	65.26
MLP + Random + NN	N=2	69.89	65.57	63.70	64.62
LR + Random + NN	N=3	68.97	64.91	61.21	63.00
SVM + Random + NN	N=3	70.35	67.46	60.50	63.79
KNN + Random + NN	N=3	66.66	59.82	69.40	64.25
MLP + Random + NN	N=3	69.65	62.04	66.41	64.15
LR + Uncertainty	N=2	79.42	75.79	76.87	76.33
SVM + Uncertainty	N=2	82.64	82.81	75.44	78.96
KNN + Uncertainty	N=2	69.89	62.25	76.87	68.79
MLP + Uncertainty	N=2	82.18	79.57	79.04	79.29
LR + Uncertainty	N=3	77.58	74.19	73.67	73.93
SVM + Uncertainty	N=3	82.49	79.93	79.36	79.64
KNN + Uncertainty	N=3	69.74	62.88	72.95	67.54
MLP + Uncertainty	N=3	81.72	77.94	77.06	77.50

Performance Measure	Seed set + random sampling + NN in the embedding space	Uncertainty sampling	SVM (n grams + embeddings)
Precision	67.17 ± 9.90%	73.65 ± 3.45%	76.49 ± 3.41%
Recall	32.35 ± 7.65%	79.39 ± 3.72%	80.30 ± 3.73%
Accuracy	82.04 ± 2.34%	75.38 ± 2.76%	77.71 ± 2.56%
F1 score	43.02 ± 7.90%	76.34 ± 2.77%	78.28 ± 2.71%
AUC	83.61 ± 2.88%	83.67 ± 2.61%	85.91 ± 2.32 %

Table 3.4: Voice-for-the-voiceless classifier performance

Chapter 4

Conclusion

In this report we provided active learning based approaches to classify comments supporting the farmers protest. As we worked with our own dataset, the results we got are based on real life scenario as it was made using comments fetched from YouTube. From the results we can conclude that we have successfully got decent accuracy in classifying comments and the methods we showed can be extended to different domains. The comparative analysis with the existing work showed that our approach in classifying comments gave similar results, sometimes even better. This work also enhanced our Socio-Political knowledge and we hope that it will also enlighten all people around the globe about this affair.

Scope for Further Research

The proposed method was able to classify comments which were fetched from YouTube only. In future we will try to fetch comments from various other sources as well as check results with more classifiers.

References

[1] Shankar, A. Indian agriculture farm acts: 2020. *International Journal of Modern Agriculture* **10**, 2907–2914 (2021).

[2] Palakodety, S., KhudaBukhsh, A. R. & Carbonell, J. G. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 454–462 (2020).

[3] Lewis, D. D. & Gale, W. A. A sequential algorithm for training text classifiers. In *SIGIR'94*, 3–12 (Springer, 1994).

[4] Zhu, X. J. Semi-supervised learning literature survey (2005).

[5] Settles, B., Craven, M. & Friedland, L. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, vol. 1 (Vancouver, CA:, 2008).

[6] Joulin, A. *et al.* Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[7] Chen, Q. *et al.* Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038* (2016).

[8] Kim, S., Kang, I. & Kwak, N. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 6586–6593 (2019).

[9] Nguyen, H. T. & Smeulders, A. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, 79 (2004).

[10] Ru, D. *et al.* Active sentence learning by adversarial uncertainty sampling in discrete space. *arXiv preprint arXiv:2004.08046* (2020).

[11] Schumann, R. & Rehbein, I. Active learning via membership query synthesis for semi-supervised sentence classification. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 472–481 (2019).

[12] Ren, Y., Wang, B., Zhang, J. & Chang, Y. Adversarial active learning based heterogeneous graph neural network for fake news detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, 452–461 (IEEE, 2020).

[13] Xiang, G., Fan, B., Wang, L., Hong, J. & Rose, C. 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*, 1980–1984 (2012).

[14] Watanabe, H., Bouazizi, M. & Ohtsuki, T. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access* **6**, 13825–13835 (2018).

[15] Scheffer, T., Decomain, C. & Wrobel, S. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, 309–318 (Springer, 2001).