

Movie Dialog Extraction and Analysis

Bhalakumaaran E R, Gopikrishnan N S, Praveen Ahuja

Computer and Information Sciences and Engineering (CISE)

INTRODUCTION

Emotions are an important aspect in human interaction. A successful computer human interaction system should be able to recognize, interpret and process human emotions.

In this project, we have performed classification of anger, disgust, fear, joy and sadness in text on the Cornell Movie Dialog dataset. We have used Support Vector Machine which focuses on classification of emotions in text.

An overall accuracy of approximately 63% for five class emotional text classification was achieved while using SVM (Support Vector Machine) classifier. We have tested and discussed the results of classification using cross-validation techniques for emotion classification and also provided powerful visualizations of the results of our statistical analysis.

OBJECTIVES

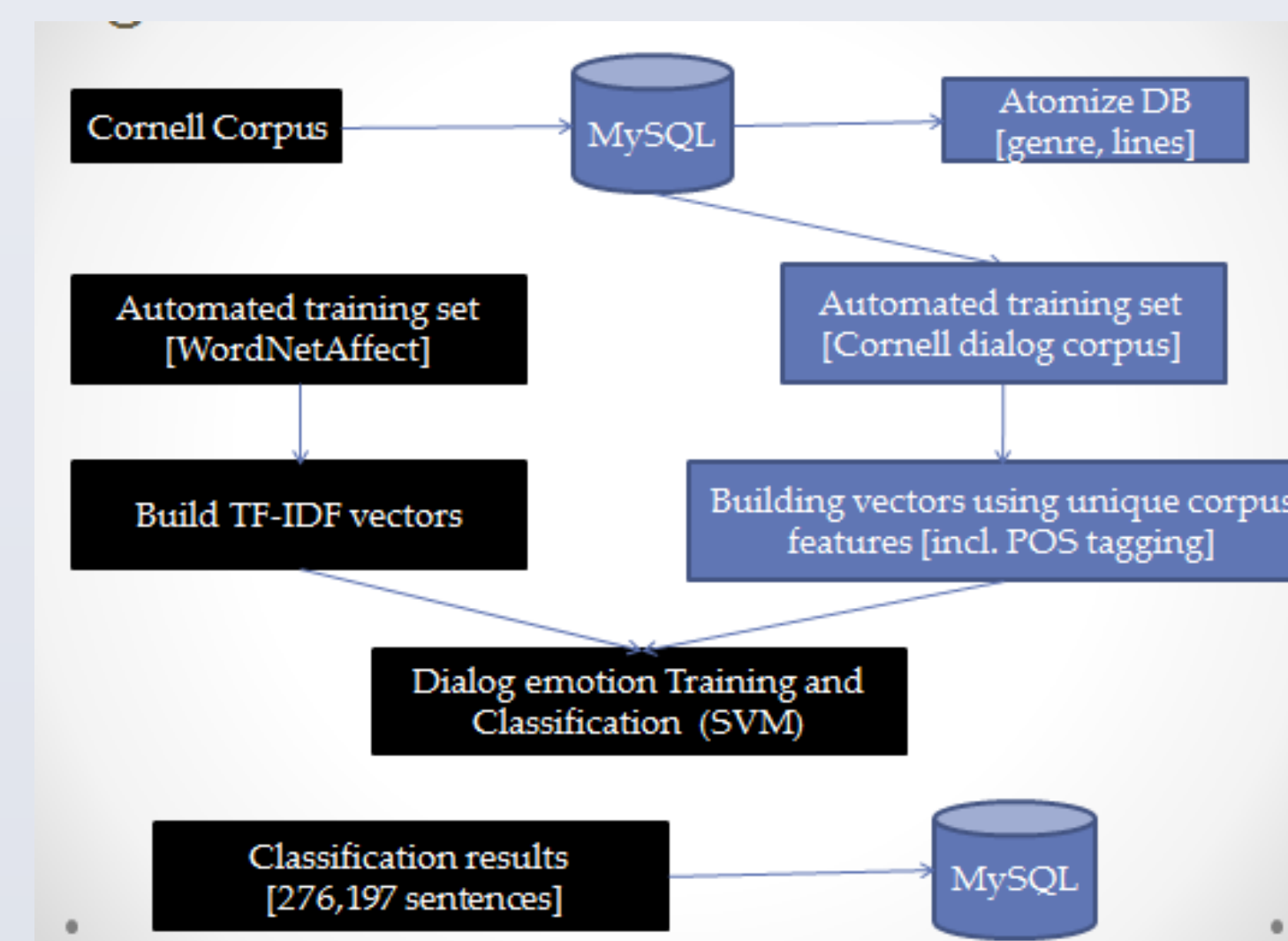
- Perform emotion classification on Cornell movie dialog corpus
- Perform data analysis on this result.
- Analyze mix of emotions across movie scripts and perform the following predictions
 - Character Analysis : Determine similar characters in different movies based on emotional content of their dialogs
 - Movie Trend Analysis : Determine the successful combination of genres across many years
 - Psychological Analysis : Determine words that best reflect the psychology behind the character based on emotion content of the dialog.

DATASETS & ALGORITHMS

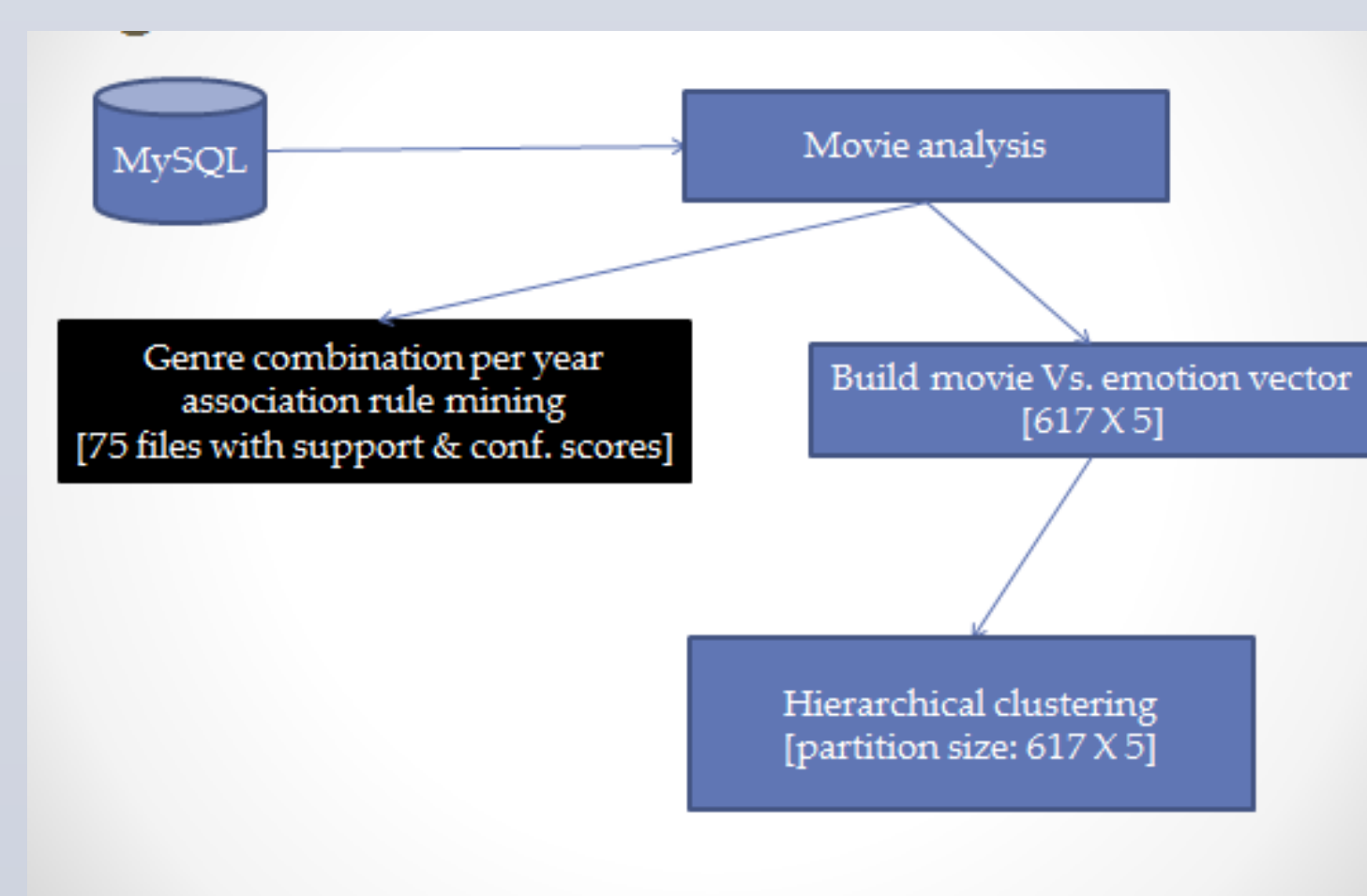
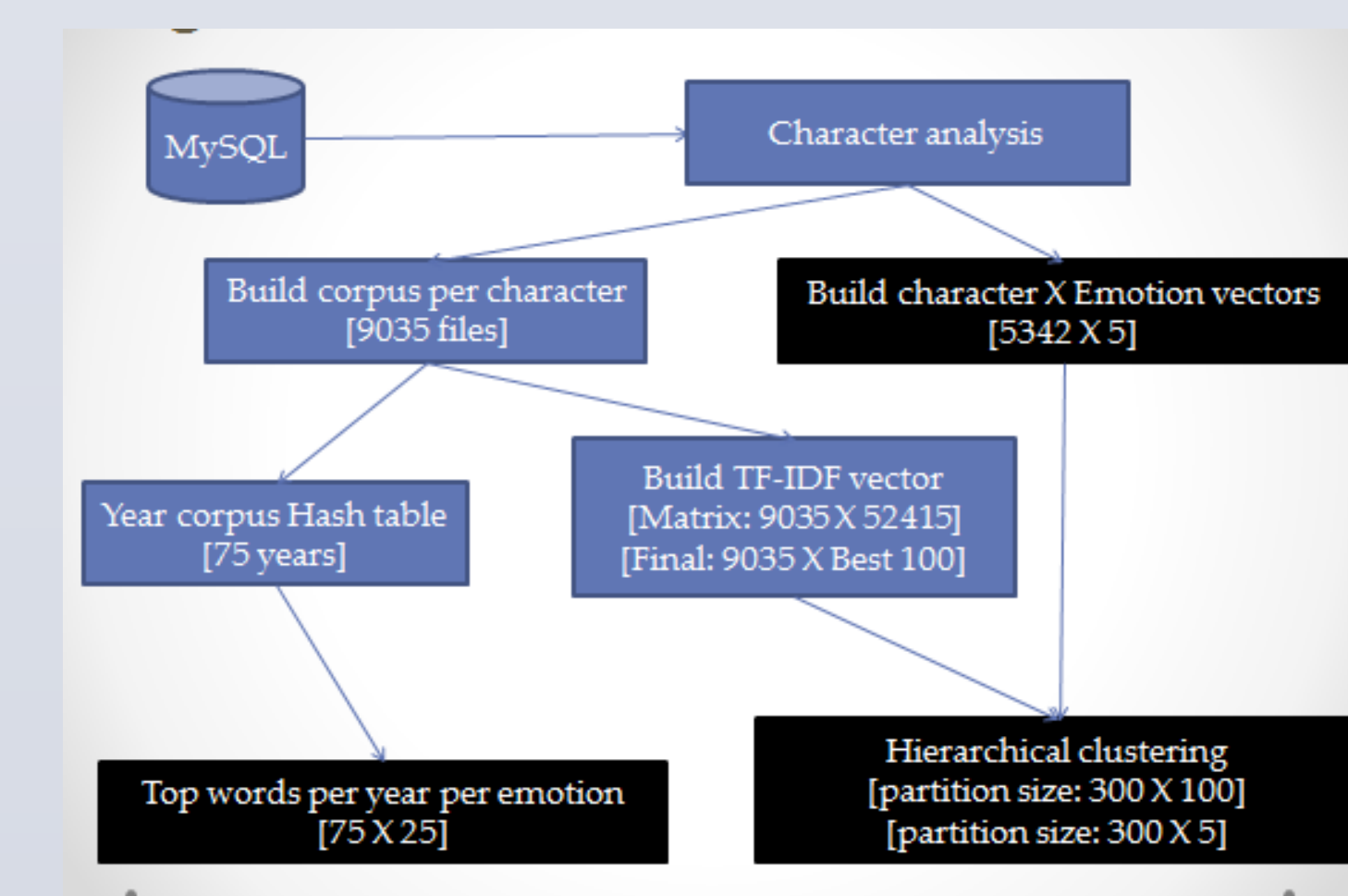
- Datasets:
 - Wordnet – Affect: An Affective Extension of Word Net [Carlo et.al., LREC 2004]
 - Cornell movie-dialogs corpus:
 - Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs [Cristian et.al., CMCL '11]
 - ISEAR dataset:
 - International Survey on Emotion Antecedents and Reactions
Labeled Sentences : 7666

Algorithm:

Phase 1



Phase 2



RESULTS

- Following are the classification results of the given dialog corpus
- The classification is based on 5 emotions namely:
 - Anger
 - Disgust
 - Fear
 - Joy
 - Sadness

	Predicted Labels				
	Anger	Disgust	Fear	Joy	Sadness
Actual labels					
Anger	776	56	115	37	101
Disgust	271	515	160	61	78
Fear	136	41	795	53	60
Joy	184	30	94	687	90
Sadness	177	40	102	115	651

Classification Accuracy:

- # labels = 5
- # sentences per label = 1085
- No. of unique tokens/feature set size = 8937
- Accuracy = 63.1152% (3424/5425)

Visualizations:

1. Movie – Genre rating:

- A comparison by means of bar charts to represent top genre combinations against the average movie ratings in that year indicating success of that combination.

2. Top emotion words:

- A tree representation of top k words used to express an emotion in that year

3. Movie versus Emotion:

- Indented Tree representing similarity between movies in a partition on mix of emotions based on Pearson Scores calculated using Hierarchical Clustering algorithm

4. Character Analysis:

- Similarity between 2 characters in/across movies based on
 - Tf – Idf (Term frequency – Inverse document frequency)
 - Emotional dialogs

CONCLUSIONS

- In this project, we have implemented a SVM classifier for emotion classification of text.
- Manual labeling of large data set for evolution of a training set is inefficient and time consuming.
- Sentences similar to corpus dialogues need to be the major part of training set documents.
- Algorithms for improving accuracy:
 - LDA/LSI (Latent Dirichlet Allocation/Latent Semantic Indexing)
 - Stop words removal
 - Stemming
 - Data Cleaning

REFERENCES

- Paper References:
 - Feeler** : Emotion Classification of Text Using Vector Space Model [Taner et.al., AISB 2008]
 - Survey** on Multiclass Classification Methods [Mohamed Aly, Technical Report, Caltech, USA, 2005]
- Existing Technologies:
 - Google prediction API
 - Mahout
 - libSVM
 - Classifier4J
 - Weka
 - Rapidminer
 - ConceptNet

ACKNOWLEDGEMENTS

- We are thankful to our instructor Prof. Dr. Daisy Zhe Wang for giving us the constant support and encouragement during the course of this project.

Instructor

- Prof. Dr. Daisy Zhe Wang (daisyw@cise.ufl.edu)

Students:

- Bhalakumaaran ER (ber@cise.ufl.edu)
- Gopikrishnan NS (gnks@cise.ufl.edu)
- Praveen Ahuja (pahuja@cise.ufl.edu)