AUTOMATED CLASSIFICATION OF EMOTIONS USING SONG LYRICS

Rajitha Schellenberg

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

December 2012

APPROVED:

Rada Mihalcea, Major Professor
Paul Tarau, Committee Member
Cornelia Caragea, Committee Member
Barrett Bryant, Chair of the Department of Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of Engineering
Mark Wardell, Dean of the Toulouse Graduate School

Schellenberg, Rajitha. *Automated Classification of Emotions Using Song Lyrics*. Master of Science (Computer Science), December 2012, 40 pp., 3 tables, 13 figures, references, 19 titles.

This thesis explores the classification of emotions in song lyrics, using automatic approaches applied to a novel corpus of 100 popular songs. I use crowd sourcing via Amazon Mechanical Turk to collect line-level emotions annotations for this collection of song lyrics. I then build classifiers that rely on textual features to automatically identify the presence of one or more of the following six Ekman emotions: anger, disgust, fear, joy, sadness and surprise. I compare different classification systems and evaluate the performance of the automatic systems against the manual annotations. I also introduce a system that uses data collected from the social network Twitter. I use the Twitter API to collect a large corpus of tweets manually labeled by their authors for one of the six emotions of interest. I then compare the classification of emotions obtained when training on data automatically collected from Twitter versus data obtained through crowd sourced annotations.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1. Emotions

Emotions are a significant aspect of human speech. In spite of various diverse cultures and languages, there are emotions expressed in various forms. Be it language, facial gestures, music or dance; emotions form an important element to convey the feeling. Emotion analysis has been an equal field of interest in both computational linguistics and psychology domains.

Emotion analysis has been a growing research field in computational linguistics. There are many areas of applications like emotional analysis in text [1], emotional analysis in music [12] [2] [3] [4] [5], mood classification in blogs [13], emotional analysis in social networks [14] etc.

This thesis presents the analysis of emotions in song lyrics by considering the textual features.

1.2. Problem Definition

Automated classification of emotions in song lyrics is the process of categorizing the song lyrics into emotions. A song can be categorized by some known features like artist name or album name, but most of the music information retrieval systems lack in emotion categorization. There are some exceptions like “allmusic.com” where content-based music categorization is done [4].

In this thesis, I try to determine whether song lyrics can be classified into various emotion categories.

1.3. Proposed Solution

I consider 100 popular song lyrics for the experiments. I use the six basic emotions; anger, disgust, fear, joy, sadness and surprise defined by Ekman [7]. The emotions for these songs are collected using Amazon Mechanical Turk [16]. The Mechanical Turk results are used as gold standard to compare and analyze the results obtained with the trained-classifier results.

I build three different classifier systems to train 90 song lyrics and test on the remaining 10 songs. I compare the classifier results with the Mechanical Turk results for analysis. I have another system where Twitter data collected for each emotion is used as the training data and tested on the 10 songs. I then combine the Twitter data and the 90 songs data to train the classifier and try to determine whether this can improve the results.

I believe that the results can be improved by enhancing the classifiers with more features and demonstrate that our systems are able to classify the song lyrics into emotion categories.

1.4. Thesis Goals

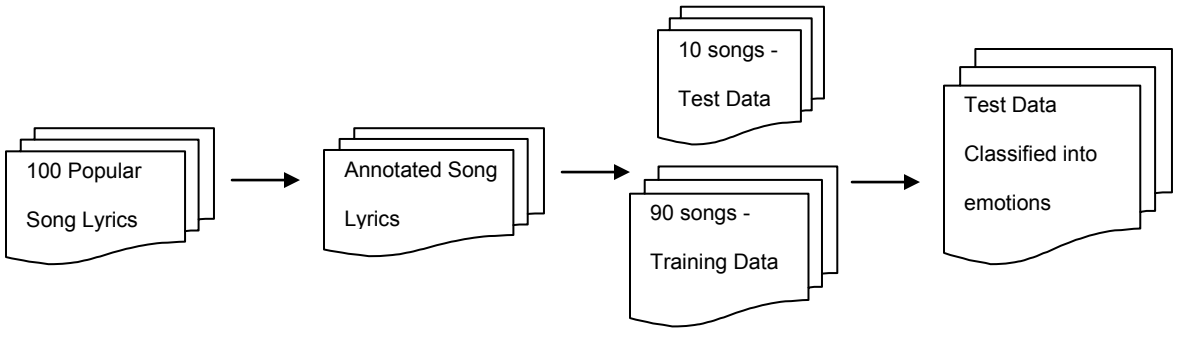


Figure 1: Goals of Classification of song lyrics.

The above figure gives an overview of different transitions of the process of classifying song lyrics into emotions. Through Amazon Mechanical Turk, I collect emotion annotations for the entire corpus of 100 songs. I then train 90 songs using the classifiers and test them on 10 songs. I plan to achieve automated classification of the test song lyrics into emotions. I can verify this by comparing them with the AMT annotations.

1.5. Thesis Outline

In Chapter 2, necessary background information and related work is outlined. The work related to identifying emotions in text and work related to analyzing emotions using both audio and lyrics is presented. Previous research in automated classification of song lyrics is briefly introduced. Research related to music information retrieval and research involving social networking data as the data set is also presented.

In Chapter 3, corpus description and collecting annotations from Amazon Mechanical Turk is explained. The Amazon Mechanical Turk is an internet platform that enables computer programmers to utilize human intelligence, through workers, to perform tasks that computers are currently unable to do [17]. The requestors post the tasks to be worked on by the people (workers) for monetary reward. In this chapter, the collection of the data set for the experiments is explained.

In Chapter 4, methods used to build classifiers are explained. Five different systems are described. The first two classifiers use the song lyrics as the dataset where a part of the dataset is used to train classifiers and tested on another part. The classifier results will then be compared to the Mechanical Turk workers results for further

evaluation. The third classifier uses naïve bayes classification method to train part of data and then test on another part. The fourth classifier uses Twitter data as the data set and is tested on the song lyrics. In this system, I use Mechanical Turk results only for the evaluation purposes and use Twitter data to train the classifier. The fifth system combines the training data set to be Twitter data and Mechanical Turk 90 song lyrics. It then evaluates the performance on the test songs.

In Chapter 5, results of these classifiers and the gold standard Mechanical Turk results are compared and evaluated further to present an interesting study of automated classification of song lyrics.

In Chapter 6, discussion on the results obtained is presented

CHAPTER 2

RELATED WORK

2.1 Background

The computational research in music emotions is a growing field. The availability of large online music databases created by vendors and the need to organize it is the motivation of this research field. The importance of music organization, classification and retrieval has been an interest to both arts and computer science communities.

2.2. Related Work

Though classification of emotions in music is still a growing field, there is a considerable amount of contribution in emotion identification in text. Emotion identification can be viewed as classification in text, classification in music that is audio and a combination of both as well.

In research related to emotion identification of text, Carlo Strapparava and Rada Mihalcea [1] presented the analysis and process of automatic identification of emotions through knowledge based and corpus based methods. The emotions taken into consideration are anger, disgust, fear, joy, sadness and surprise defined by Ekman [7]. The importance of emotions in various fields like psychology, behavior sciences and computational linguistics has been highlighted in this paper. Some of the applicative scenarios mentioned are sentimental analysis, computer assisted creativity and verbal expressivity in human computer interaction.

The dataset considered in these experiments are news headlines from various newspapers and the primary reason for this choice being news headlines is that they

have more emotional content. For each headline, annotators were asked to annotate them on a scale of the above mentioned six emotions to make it an unsupervised setting. The inter-annotator agreement was made using Pearson correlation. Two methods of evaluations were made. Fine-grained values were taken into consideration through Pearson correlation and coarse-grained evaluation values were 0 and 1 for a particular emotion. The latter was analyzed using precision, recall and F-measure.

The knowledge based emotion analysis mentioned in the paper; the experiment used Word Net AFFECT which is an extension of Word Net database to classify the direct affective words. An algorithm was used to determine the frequency of words present in the news headlines to compute a score that was then used for further analysis. Latent semantic analysis is a vector space model representation by considering tf-idf weighting. In this paper, it is mentioned about the experiments conducted using various representations like (i) the vector of the specific word denoting the emotion (e.g. “anger), (ii) the vector representing the synset of the emotion (e.g. {anger, choler, ire}), and (iii) the vector of all the words in the synsets labeled with the emotion. However the synset is built, similarity measure can be determined and evaluated.

The corpus based emotion analysis used the blog posts from LiveJournal.com as a corpus. The paper says that every blog community practices a genre of writing; this particular community has topics related to events of everyday life. After cleaning up the blog data, they trained a naive bayes classifier considering that particular emotion blogs as positive examples and all other emotion blogs as negative examples.

After analyzing the data required by both knowledge based emotion analysis and corpus based emotion analysis, the experiments were conducted by the following systems.

1. WN-Affect presence, which is used as a baseline system, and which annotates the emotions in a text simply based on the presence of words from the Word-Net Affect lexicon.
2. LSA single word, which calculates the LSA similarity between the given text and each emotion, where an emotion is represented as the vector of the specific word denoting the emotion
3. LSA emotion synset, where in addition to the word denoting an emotion, it's synonyms from the Word Net synset are used as well.
4. LSA all emotion words, which augments the previous set by adding the words in all the synsets labeled with a given emotion, as found in Word Net Affect.
5. NB trained on blogs, which is a naive bayes classifier trained on the blog data annotated for emotions.

As expected, different systems have different strengths and hence the results. The experiments mentioned in this paper give us the insight to determine what works better for a particular emotion or method. These experiments also give a lot of scope to explore the semantic details of emotions by enhancing these methods.

A complete review of existing work in music emotion identification developed in psychology and engineering has been presented by Yi-Hsuan Yang and Homer H. Chen [2] [3]. They introduced emotion based music retrieval and organization methods and ranking-base emotion annotation. Their work includes a combination of information extracted from lyrics, chord sequence and genre metadata for better accuracy.

Another related work closely associated to the previous work is by Dan Yang and Won Lee [4]. In this paper, machine learning techniques are used instead of human experts to extract emotions in music. To identify emotional value in music information

retrieval, more than human expertise is needed. The author says that human expertise based on fine arts or psychology takes a long time to achieve, about one PhD per artist, and faster solutions need to be found. Building a machine learning system with large quantity of online music would offer a faster, automated solution.

The author tells us that emotion identification requires a better understanding of the text along with its structure and the context used. Statistical natural language processing techniques hinders the common sense intuition of emotions. The analysis of existing systems in text mining can be studied as psychological feature identification in text and linguistic research.

In analysis of psychological feature identification, psychologists interpret the affective value of words based on empirical surveys and expert judgments. Number of words were rated basing on dimensions such as valence, intensity, dominance and sometimes psychological categories as well.

Linguistic research on psychological issues explains about how the structure and rule-like nature of language provides clues to distinguish emotion content in sentences. The author provides description of many text-retrieval, sentiment identification and emotion polarity systems.

The experimental design described in the paper takes analysis of song lyrics data into consideration. The number of unique words did not even increase according to the number of documents being added following the Zipf's law. Each additional song covers very few new words making it hard to collect a sufficient number of songs. To solve this problem of feature sparseness, a number of feature reduction methods commonly used in text mining are more suitable to topic identification rather than

emotion identification. The 168 emotion categories in allmusic.com where generalized to 23 emotion categories in this paper. The system was built with 23-class classification on 1032 songs with over 5000 unique words.

The author says that specific discrete emotions are difficult to distinguish using acoustic data alone. The results for positive and negative emotion identification in lyrics reflect show the similar success rate, differently from acoustic feature extraction where the negative emotions are difficult to distinguish on the basis of acoustic features alone. This approach of considering emotion in music in terms of the composition of both acoustic and textual effects corresponds to different levels of emotion processing in the brain.

This paper proposed a new approach in identifying discrete emotions instead of just polarity and used feature vectors instead of bag-of-words on song lyrics. Their results for mining the lyrical text based on emotions are promising, generate classification models that are humanly-comprehensible, and generate results that correspond to commonsense intuitions about specific emotions.

Another work which considered acoustic information for audio analysis and semantic information for text classification is by Laurier C, Grivolla. J & Herrera P [5]. In this paper, approaches to classify songs in different mood categories using both audio based techniques and using semantic information from song lyrics are well presented. This paper is focused to study the complementarity of lyrics and audio and also an integrated approach.

The database is a collection of 1000 songs which fall into 4 categories that is happy, sad, angry and relaxed. Happy and relaxed with positive valence and high and

low arousal respectively; “angry” and “sad” with negative valence and high and low arousal respectively. They also considered a complimentary categories being not happy, not angry etc.

To classify basing on audio, acoustical information and descriptors derived from state-of-the-art research in music information retrieval are being used. Accuracies for the four categories are determined using SVM, logistic and randforest algorithms. The performance of SVM was better and accuracies are above 80% for each of the categories.

In lyrics classification, the first approach was based on similarity using Lucene. In this, the representation of songs is reduced to a bag of words, and then used in a Lucene document retrieval system to rank documents by their similarity. Tf-idf approach is used and also an optimal number of documents are considered for better performance. This method fetched an average accuracy around 60% and was difficult to be integrated with the audio-based approach.

In the next approach, LSA, lyrics are projected into a lower-dimensional space and are used in combination with tf-idf weighting. The use of LSA doesn’t dramatically improve performance. The third approach is based on language modeling, where terms with a large relative difference in document frequency in one class being multiple times that in the other class are considered. This approach used with SVM classifier performed the best and very close to audio based approach.

Combining all the features of lyrics and audio in the same space within one classifier outperformed yielding better results than any other approaches.

Other related work includes music information retrieval (MIR). MIR is like any other information retrieval system where music or songs or information related is retrieved. MIR is a small but growing field of research with many real-world applications. MIR is usually one or more combinations of the following domains, musicology, psychology, academic music study, signal processing or machine learning. The quantity of tracks is ever increasing and posing challenges for MIR system to classify and categorize them. Music genre categorization is a common task for MIR. There are machine learning techniques applied to automatically classify music as it is inherently expressive of affective meaning.

Closely related to this is the work presented by Lijuan Zhou, Hongfei Lin, and Wenfei Liu [6]. The authors describe the implicit emotional association between users and music to enrich MIR. The latent emotional intent of queries is studied via machine learning based emotion classification and compared the performance of emotion detection approaches on different feature sets.

CHAPTER 3

CORPUS DESCRIPTION

3.1. Lyrics of Song

To enable our experiments I have considered a data set of 100 popular songs (like “bad romance” by Lady Gaga, “hotel California” by Eagles, “so this is Christmas” by Celine Dion etc.). Music can affect our emotions; makes us feel happy, makes us cry, makes us dance. Though, just the audio can impact our emotions sometimes, I consider only the lyrics of songs for our experiments and study.

To explore our classification experiments using emotion annotations, I need a gold standard to compare with the emotions derived. The gold standard emotion annotations for 100 song lyrics are determined by using Amazon Mechanical Turk. The process of collecting these manual annotations using Amazon Mechanical Turk is explained in the following sections.

3.2. Annotations of Songs

3.2.1. Introduction to Amazon Mechanical Turk

Amazon Mechanical Turk (AMT) is a market place for work, where developers can request workers to work on human intelligence tasks (HITs). The person who requests to perform work on human intelligence tasks is known as a requestor. Workers can select to work from thousands of available tasks. HITs – human intelligence tasks can be like categorizing videos, identifying objects in the image, writing reviews of products etc. A Mechanical Turk worker can work from home at their convenient hours and get paid for the genuine answers. A Mechanical Turk requestor can choose their

workers based on the location or qualification of the worker. The qualifications of a worker can be his skill in a certain domain, ability or reputation. For example, requestors can choose workers based on number of HITs performed by workers in their life time, the approval rate of workers etc.

Requestors can get their work done in minutes and pay the workers only when they are satisfied with their work. Number of workers working on a HIT can be specified by the requestor. The requestor can see the status of progress of the workers. Requestors can choose to accept or reject the workers HITs which affects their approval rate.

3.2.2. Publishing HITS (Songs) in AMT

Having spoken about AMT, in this section let us see how to publish HITs, that is lyrics of songs for emotion classification. A requestor account is needed to publish HITs. Once a requestor account is created, I can create or edit existing templates from the design sections. Note that the template is just like a HTML prototype, which when once created can be used to upload many songs. The template consists of three sections entering properties, design layout and preview/finish.

In the properties section; title of the HIT, description of HIT explaining to the workers, keywords which help the workers to look up for the HIT, time allotted for a worker to complete a HIT, HIT expiration details, worker qualifications, workers payment details can be mentioned. I can choose how many workers can work on a particular HIT. I have chosen 10 workers to work on a single HIT/song.

In the design layout, HTML source can be edited according to the requirements. In the HTML source, instructions to the worker can be explained. In the present scenario, HTML source contains instructions to the worker to identify emotions felt by the writer in the lyrics of a song. Workers are asked to give a score (annotation) on a scale of 0 to 10, to classify each sentence of a song into one of the six emotions: anger, disgust, fear, joy, sadness and surprise. The HTML template has each sentence with six emotions along with a text box for the workers to enter their score. Workers can enter scores for one or multiple or none of the emotions for a sentence. Some examples are also provided to the users.

After the template is created, I can preview how it looks to the worker. Next step is to upload input data into the selected template. The lyrics of a song need to be in the CSV file format and are uploaded into the template to create HITs. Now the actual HIT can be previewed where the instructions as well as sentences of the song uploaded with emotion – text boxes are present. The final confirmation to publish requires the requestor to fund their account enough to pay for the number of workers assigned to perform the task.

3.2.3. Avoiding Spam Workers

There might be annotations given to the lyrics randomly without reading and understanding the context by few workers. To avoid such spammers, there should be a way to eliminate such annotations. The lyrics of a song are transformed from a text file to XML file. And the XML file is converted into a CSV file and then uploaded to the template created in which is explained in the previous section. Along with the lyrics

sentences; a sentence asking the workers to enter “7” in all the emotion text boxes is inserted at a random point. I can capture the workers who do not enter 7 to this sentence. This allows us to eliminate the spammers and their work.

The following figure shows us a pictorial representation of the pre-processing steps involved in collecting annotations.

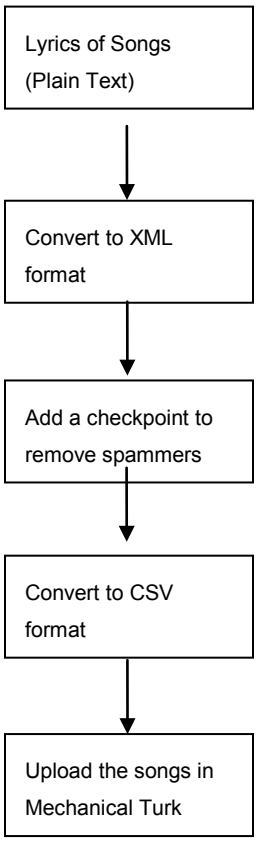


Figure 2: Workflow of pre-processing steps.

3.3 Post-Processing

3.3.1 Collecting Output from AMT

The progress of the published HITs (collectively known as batches) can be viewed in the “manage” tab of the AMT portal. Management of batches is made easy with the details provided along with each HIT. The number of assignments completed

and the number of assignments yet to be completed is clearly indicated in the details. Once the HITs are completed, the CSV file of that HIT can be downloaded following the link.

3.3.2 Annotation Analysis

Using PERL scripts, the CSV files are processed to create a space-separated text file with all the annotations. The output will include the annotations for each emotion, for each verse, for each user. Users are listed all on one line, in pairs of [user-annotation average-of-all-users]. For example,

```
ANGER BADROMAN.9 WANT__YOUR__U-GLY__I__WANT__YOUR__DI-SE-A-SE
0 2.55 0 2.55 0 2.55 4 2.11 8 1.66 2 2.33 7 1.77 2 2.33 0 2.55 0 2.55 2.29 1.75
```

Figure 3: Sentence with user annotations and their average.

This is the annotation for the “anger” emotion for a verse; first user entered 0, the average of all but the first user is 2.55 etc.

The checkpoint line inserted at a random point should have “7” for each of the users, if they are not spammers. If the checkpoint line doesn’t have 7 for any of the user, that user’s work can be eliminated considering the fact that the user has not read and understood the context while annotating. For example, in the following sentence, users 1, 3, 5 and 9 have entered a value other than 7 and should be eliminated.

```
ANGER CHECKPOINT PLEASE__ENTER__7__FOR__EACH__OF__THE__SIX__EMOTIONS
0 5.55 7 4.77 0 5.55 7 4.77 8 4.66 7 4.77 7 4.77 7 4.77 0 5.55 7 4.77 5 2.04
```

Figure 4: Check point sentence to eliminate spammers.

After eliminating the spammers, the genuine annotations are collected for the next processing step. A script is used which determines a comparison among workers

which is explained in the next section.

3.3.3. Pearson-Correlation

I compare the annotation of the users to fetch inter-annotator agreement values. The output file of the previous step, which has sentences—user-annotation—average-of-all-users is used as an input to a script which determines the outliers using Pearson correlation [18].

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment" that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name [5].

Pearson correlation is a number between 0 and 1 that measures the degree of association between two variables (call them X and Y). A higher value for the correlation implies a stronger association (large values of X tend to be associated with large values of Y and small values of X tend to be associated with small values of Y).

Using Pearson correlation gives us annotator agreement values between 0 and 1. I remove the outliers (annotators with small agreements, i.e., below .50).This leaves out only the workers whose inter-annotator agreement is good enough to consider for further experiments.

3.4. Example of Annotation

After selecting workers with good annotator agreement, a script is used to get the

aggregate value of all workers for each verse and for each emotion. The following is an example of verse from the song, “heart of gold.”

<sentence id="HEARTOFG.4">
FOR A HEART OF GOLD
<emotion category="anger" intensity="0.28" />
<emotion category="fear" intensity="0.14" />
<emotion category="joy" intensity="5" />
<emotion category="sadness" intensity="1.42" />
<emotion category="surprise" intensity="0.85" />
</sentence>

Figure 5: Integrated emotion values obtained after post-processing.

The following figure shows a pictorial representation of the post-processing steps involved in collection emotion annotations.

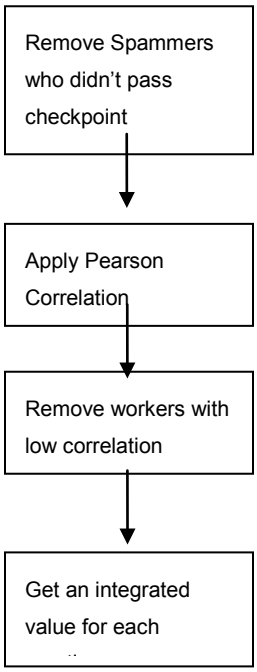


Figure 6: Representation of post-processing steps.

CHAPTER 4

METHODOLOGIES

There have been discussions about many popular classification methods like naïve bayes [9], support vector method machines and kNN [8]. Mandel and Ellis used SVM to automatically identify the artist of the song based on features calculated over their entire lengths [10].

In this paper, I use different classification methods in the experiments which are explained in the relevant sections to analyze the emotion content in song lyrics. I use textual features from song lyrics for determining the emotion content in a line.

4.1 Bag of Words

I use a bag-of-words representation of the lyrics to derive unigram counts, which are then used as input features. Firstly, I build a vocabulary consisting of all the words, including stop words, occurring in the lyrics of the training set. I do not remove the tokenization as well, as hyphens mark syllable boundaries in song lyrics [11]. Later, a comparison is made by removing the stop words and by removing tokenization as well.

After cleaning up spammers and less correlation workers (explained in post-processing section of corpus description chapter), I have resultant annotations for each sentence in a song. So, I have the entire training data (90 songs) divided into 6 bags of words basing on 6 different emotions. Each bag of words, contain words and the human-annotated value of that emotion multiplied by the number of times that word appears in that sentence.

For example, the following sentence, “love love love, I want your love” has emotion annotations as anger – 1, disgust – 0.5, joy – 8 and sadness – 2. In this sentence, the word “love” appears to be four times when compared to words like “want” or “your” or “I.”

```
<sentence id="BADROMAN.12">
LOVE LOVE LOVE I WANT YOUR LO-VE
<emotion category="anger" intensity="1" />
<emotion category="disgust" intensity="0.5" />
<emotion category="joy" intensity="8" />
<emotion category="sadness" intensity="2" />
</sentence>
```

Figure 7: An annotated sentence used for bag of words for each emotion.

By considering the number of times a word appears in the sentence and human-annotated value as well, each word is given a value based on its contribution to a given emotion to the sentence. So I consider that the word “love” has more contribution towards the emotion content and consider this word’s emotion annotations as anger – 4, disgust – 2, joy – 32 and sadness – 8 whereas emotion annotations of other words (I, want, your) are anger – 1, disgust – 0.5, joy – 8 and sadness – 2.

The following are the methods used to classify the training data.

4.2. Classification Using Bag of Words

4.2.1 Considering Product of the Words in a Sentence

This classification method checks for the presence of the test sentence’s words in the bag of words for each emotion. For each emotion, the word’s value in that emotion is taken into account which is normalized by the frequency of the word in the entire training corpus. For a given sentence, the product of each word’s value is considered for each emotion. So, I have a certain value for each emotion. If a word is

not present in the given bag of words, a very small value is assumed to be the word's value; this implies that the value of the total product is lowered. If a word is not contained in the bag of words, the entire sentence is given less weightage for that emotion. After classification, each sentence of the test songs has six values for each of the emotions.

For example, consider the following sentence, "I-'VE BEEN SO MA-NY PLA-CES." For a particular emotion, say, "anger" I check for the existence of the words "I-'VE," "BEEN," "SO," "MA-NY" and "PLA-CES" in anger – bag of words. If the words are present, there should be a value for that particular word which is then normalized by the frequency of the word in the whole training corpus. The normalization is done to eliminate the importance of common words like "the," "an," "and" etc. The product of all the scores for all the words in the sentence is considered. If a word is not present in anger – bag of words, a very small value like 0.001 is assumed to be its value. This lowers the entire product value giving the whole sentence less weightage.

This gives each sentence a value for the six emotions. This can be compared with the human-annotated value to analyze the performance, as explained in the chapter 5.

4.2.2. Considering the Sum of the Words in a Sentence

This classification method also uses the similar bag of words for each emotion. It checks for the presence of the test sentence words in the bag of words for each emotion. For each emotion, all the word's values are added and normalized using the square root of sum of the squares of all the words in that emotion.

If a word is not present in the given emotion, a very small value is assumed to be the word's value. However, it doesn't change the sentence value much. After classification, even this method will have six values for each of the emotions for each sentence of the test songs.

For example, consider the following sentence, "I-'VE BEEN SO MA-NY PLA-CES." For a particular emotion, say, "anger," I check for the existence of the words "I-'VE," "BEEN," "SO," "MA-NY" and "PLA-CES" in anger – bag of words. If the words are present, there should be a value for that particular word which is then normalized by the square root of sum of squares of all the words in that sentence. The sum of all the word's value is considered. If a word is not present in anger – bag of words, a very small value like 0.001 is assumed to be its value. This doesn't lower the entire value unlike the previous method. It just doesn't give any importance to that word.

This gives each sentence a value for the six emotions. This can be compared with the human-annotated value to analyze the performance which is explained in the chapter 5.

4.3. Processing Data

The original files with song lyrics have words with special characters. The debate was whether to keep the tokenization or not, as few special characters (hyphen marks syllable boundaries) have importance in lyrics of a song.

To find answers for the questions above, initially only 90 songs are considered as the whole corpus out of the total 100 songs. The remaining 10 songs were not used until the working systems are finalized. Out of these 90 songs, 70 songs were

considered as the training data and 20 songs as the development data. The classification methods were used to train these 70 songs and tested on the development data of 20 songs.

The data set was evaluated using both the classifiers on raw bag of words, which is non-tokenized data and which has stop words in them. Then, the data set was evaluated using both the classifiers on bag of words, which is tokenized data but the stop words were still not removed. Finally, the bag of words which is tokenized and which doesn't have stop words is considered for the evaluation. The results for the different systems are explained in detail in the evaluation section.

4.4. Naive Bayes Classification

A text classifier is an automated means of determining some metadata about a document. Text classifiers are used for diverse needs such as spam filtering, suggesting categories for indexing a document created in a content management system. One such classification method - naïve bayes [9] is used for the project. The classifier used here determines which of a set of possible categories a document is most likely to fall into.

I use naïve bayes classification method to train 90 songs and test on 10 songs. I feed each sentence from the test songs to the classifier system to classify the sentence into emotion categories. I then have naïve bayes classifier computed scores for each sentence which can be compared to the Amazon Mechanical Turk annotated emotions using Pearson correlation. The evaluation is explained in the evaluation chapter.

4.5. Classification Using Twitter Data

4.5.1. Twitter Data

Twitter is growing in popularity as a network to connect people. Users share a short message called tweets that contain a maximum length of 140. Through Twitter, users can keep up with friends as well as with the most popular trends. They express their emotions about anything and everything. In this section, these tweets are collected and used for a system to predict the emotions present in song lyrics

4.5.2. Data Set

Twitter data collected for the six emotions is used as the training data to train the classifier. The test data is lyrics of 100 songs. The lyrics of songs are annotated using Amazon Mechanical Turk which has been explained in the “corpus description” section. The classified scores are evaluated across the annotated scores. Next section explains how the data has been collected and what methodologies are being used in this system.

Twitter data is collected using TAP (Twitter access via Perl) package provided by Dr. Rada Mihalcea. The TAP package includes scripts to facilitate access to the Twitter API. The emotions for which the data is collected are anger, disgust, fear, joy, sadness and surprise defined by Ekman [7]. To result in more number of tweets, I have used synonyms of the emotions as well. For example, for the emotion “joy,” synonyms like “happy,” “cheerful” are also used.

The scripts in Twitter API covers two main functionalities; firstly the keyword search and secondly monitoring of a user's activity. The scripts will result in a corpus of tweets with one of the following two properties:

(1) Tweets that match a certain query; or

(2) Tweets that are written by a given user, who has a public profile.

Each tweet in the corpus will include:

- The tweet ID - a unique ID assigned by Twitter to every tweet
- The date and time of the tweet
- The screen name of the user who posted the tweet
- The tweet itself

The scripts also have the ability to "update" a collection of tweets by keeping track of the ID of the last tweet collected in a given session and continuing the collection of tweets from that ID onward. This has the benefit of avoiding duplicates in the data collection (e.g., if the same search is run at two different points in time), and also to avoid overloading the Twitter server.

Each emotion’s data collected from twitter can be treated as a bag of words which is parsed and cleaned up. These bags of words are used to train the classifier and tested on lyrics of 100 songs. The emotion annotations retrieved through Amazon Mechanical Turk are being used as a gold-standard to verify the classification results.

4.5.3 Method

I have used naïve bayes classification method. Once the classifier is trained with six emotions, meaning the data collected from twitter, the document to be tested is passed. Each sentence of a song is passed as a document to be tested. Each sentence will have computed scores for the six emotions. Such scores are computed for all the sentences in 100 songs. For evaluation purposes, each emotions score is summed up

for a song. This results in a cumulative score for each of the six emotions for a given song. In a similar way, the annotated scores of each emotion per sentence are summed up to give cumulative annotated score for each of the six emotions for a given song. The comparison and analysis of these two scores is explained in evaluation section.

4.6. Classification Using Integrated Data Set

I combine the training data set to be Twitter data and AMT annotations for the 90 songs for the six emotions. I consider twitter corpus for all the six emotions and also include the emotion annotations fetched from AMT for the 90 songs. I train naïve bayes classifier with this integrated data set and test on the 10 songs. I have combined the data set to determine whether I could fetch best of both the worlds.

CHAPTER 5

EVALUATION AND DISCUSSION

5.1. Evaluation of Classification Using Bag of Words

Each sentence of the 10 test songs will now have classifier computed scores and the human annotated scores for all the six emotions.

Pearson correlation is applied to each sentence's scores. This means, for each sentence I have six emotion scores generated by the classifier and six emotion annotated score collected through Amazon Mechanical Turk. When Pearson correlation is applied to each sentence's scores I have values between "0" and "1." A value close to "1" shows us that the classifier generated scores are in close-agreement with annotated values.

The following data and the charts in the next section show the performance of the 10 test songs across a Pearson correlation score of 0 to 1. Figure 8 shows the performance of 10 test songs when the product of the scores is considered and Figure 9 shows the performance of 10 test songs when the sum of the scores is considered.

5.2. Evaluation of Naïve Bayes Classification

The results of this classification method are shown in Figure 10. The y-axis is for the correlation (value between 0 and 1). The x-axis determines the number of sentences in that song.

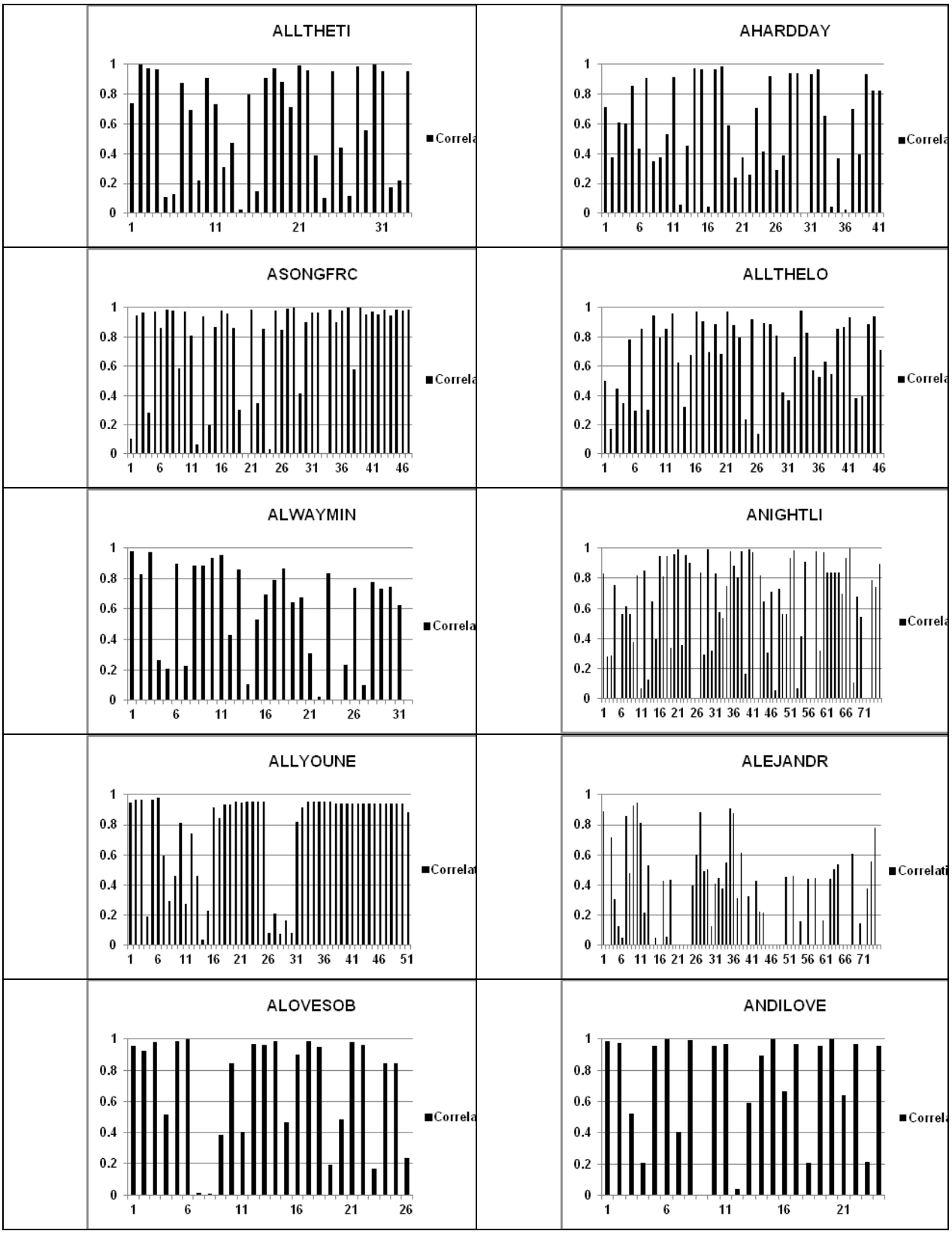


Figure 8: Classifier using Bag of Words – product of word scores in a sentence.

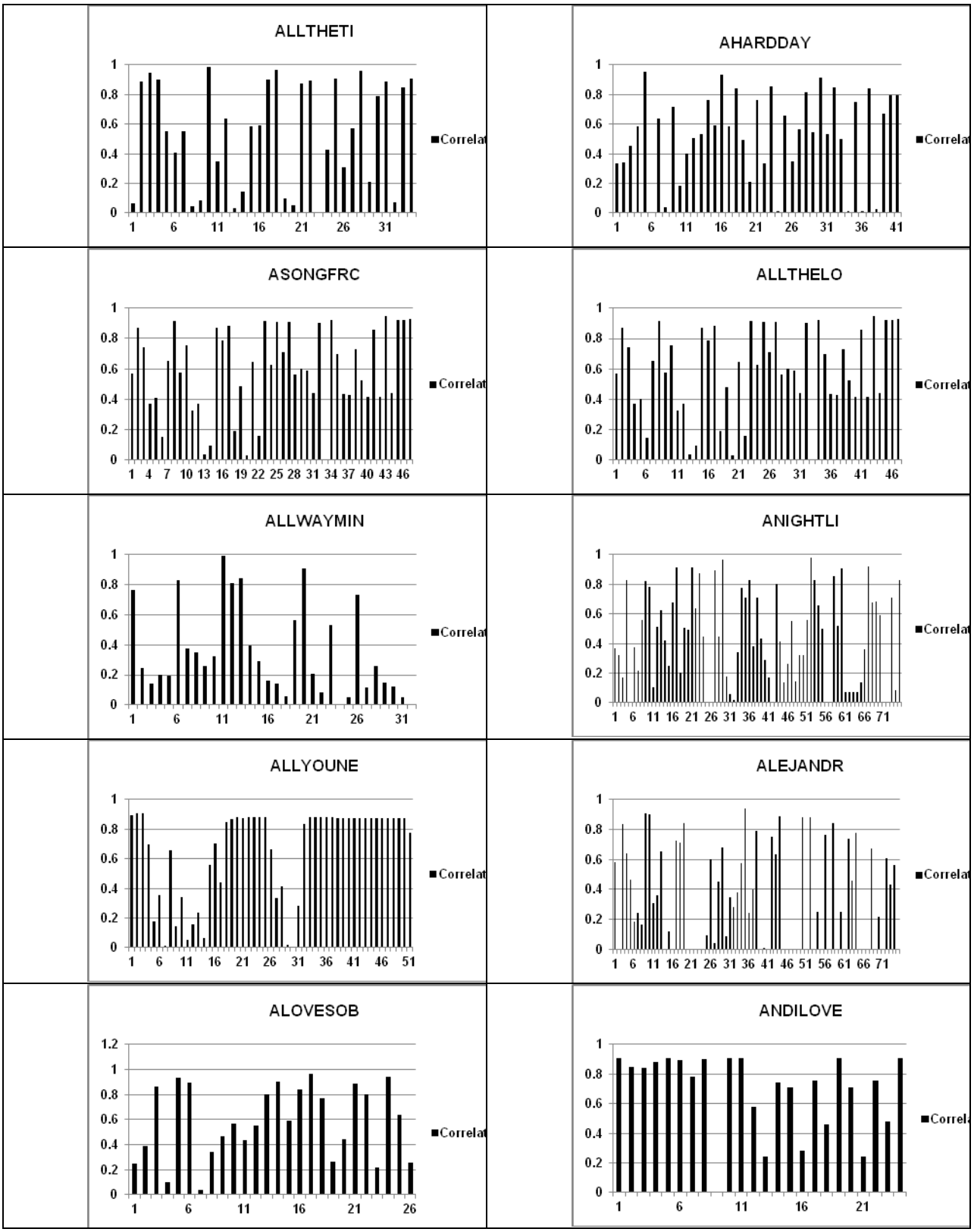


Figure 9: Classifier using Bag of Words – sum of word scores in a sentence.

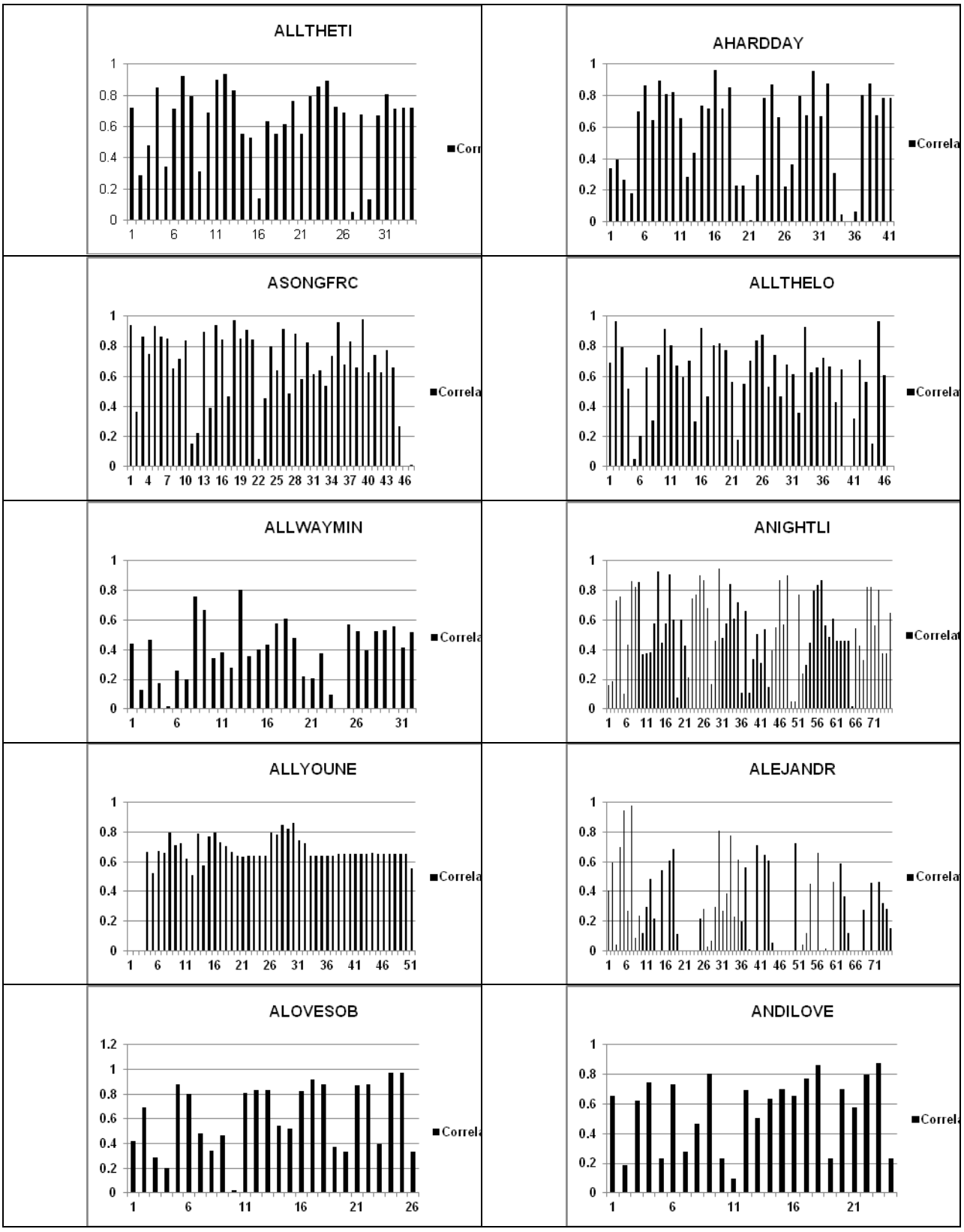


Figure 10: Evaluation of NB classifier.

I determine the performance of our systems by doing an emotion-wise analysis. I consider the six emotion scores summed up for the 10 songs and compare it with annotated scores. The values shown in the following table is the correlation values for each emotion for different systems.

Table 1 explains the performance of each emotion in different classification methods.

Table 1: Emotion Scores for Different Classification Methods

	Classification using bag of words		
Emotion	**Product of word scores**	**Sum of word scores**	**Naïve Bayes**
Anger	**0.3174**	**0.1704**	**0.1933**
Disgust	**0.2858**	**0.1434**	**0.1783**
Fear	**0.1501**	**0.0636**	**0.0812**
Joy	**0.1967**	**0.3117**	**0.1439**
Sadness	**0.1495**	**0.1113**	**0.0505**
Surprise	**0.1411**	**0.054**	**0.1685**

5.3. Evaluation of Classification Method based on Twitter Data

Twitter data is used to train the naïve bayes classifier and then tested on the songs. Each sentence of a song is passed on to the classifier as a test document and scores are computed for the six emotions. Pearson correlation is used to give a correlation value between the classifier computed scores and human annotated scores in every sentence. The correlation value for each sentence tells us how accurate the classifier computed emotion scores are when compared to the human annotated scores. The results of 10 test songs are shown in Figure 11.

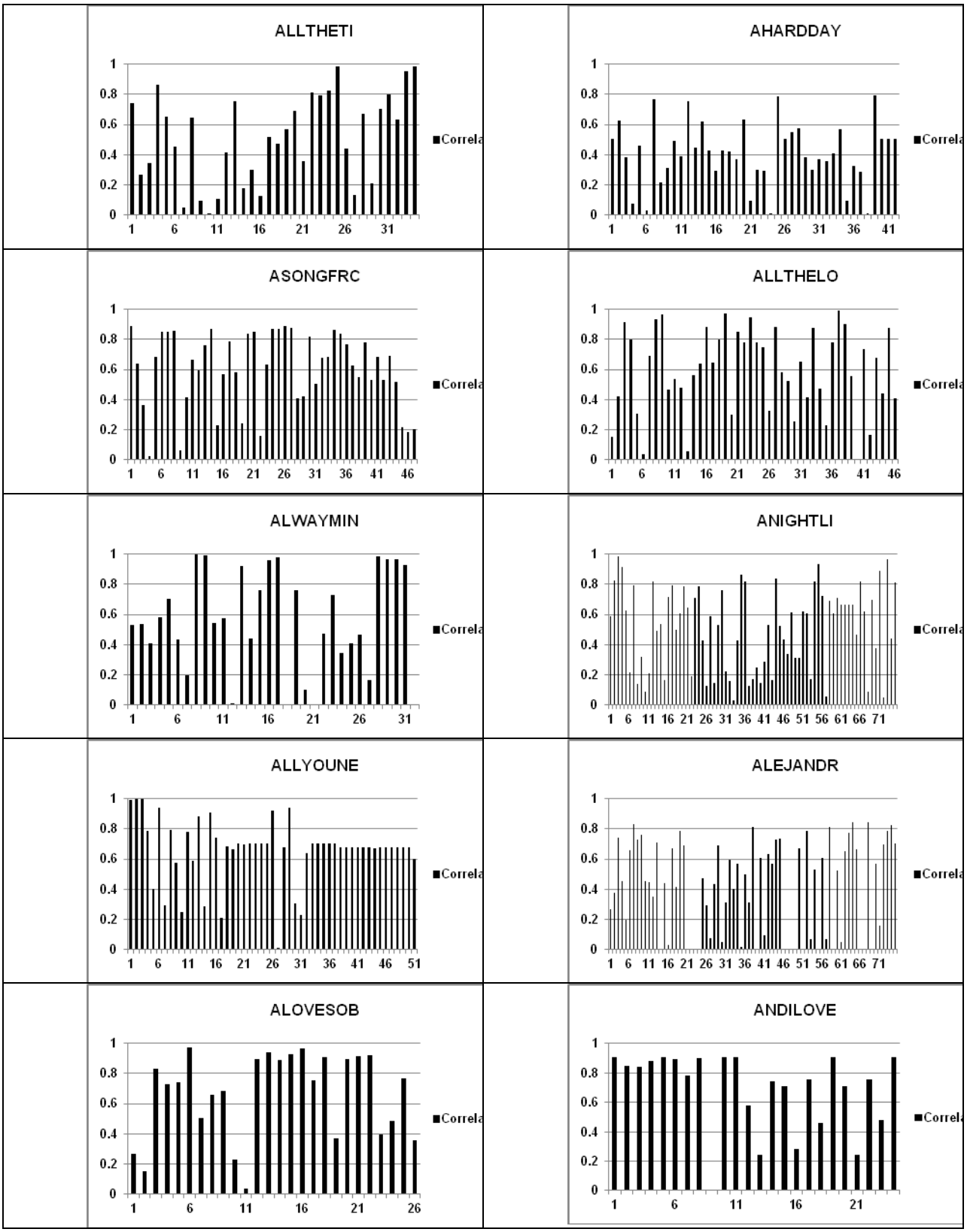


Figure 11: Evaluation using Twitter data.

I make another evaluation across the emotions using the Twitter data. This method is different from the previous methods as I train the classifier using twitter data and do emotion wise evaluation. For the six emotions, I have six values for classifier computed scores for each of the 10 songs. Pearson correlation is applied to human annotated scores (using Mechanical Turk) and classifier computed scores for each emotion to fetch us the correlation value for each emotion. Table 2 and Figure 12 show the performance of different emotions with their correlation scores.

Table 2: Correlation Values for Each Emotion when Twitter Data is Used

Twitter	Correlation
Anger	0.3956
Disgust	0.3379
Fear	0.2814
Joy	0.4927
Sadness	0.2968
Surprise	0.3535

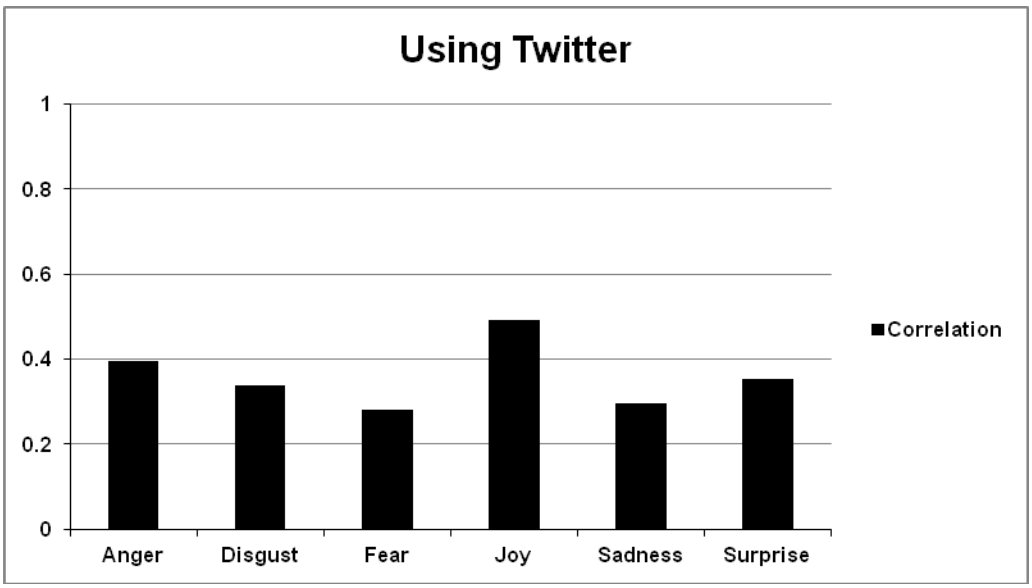


Figure 12: Comparison of correlations values for six emotions computed using 100 songs.

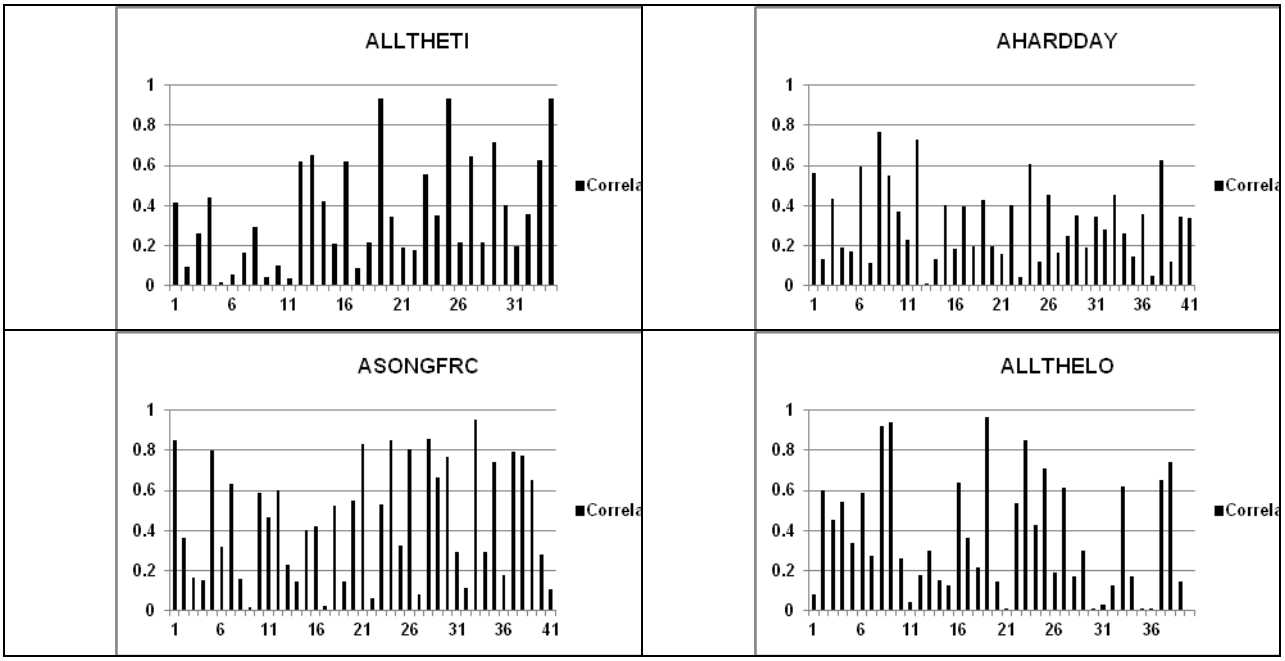
5.4. Evaluation of Classification Method based on Integrated Data

I combine the data set of Twitter data and the annotations of 90 songs to train the naïve Bayes classifier. Figure 13 shows the 10 test songs performance in this system.

Table 3 shows us the performance of emotions alone when using this integrated data set.

Table 3:Correlation Values for Each Emotion when Twitter Data and 90 Songs Data is Used

Combined Data set	Correlation
Anger	**0.3242**
Disgust	**0.2632**
Fear	**0.4836**
Joy	**0.4730**
Sadness	**0.3748**
Surprise	**0.6151**



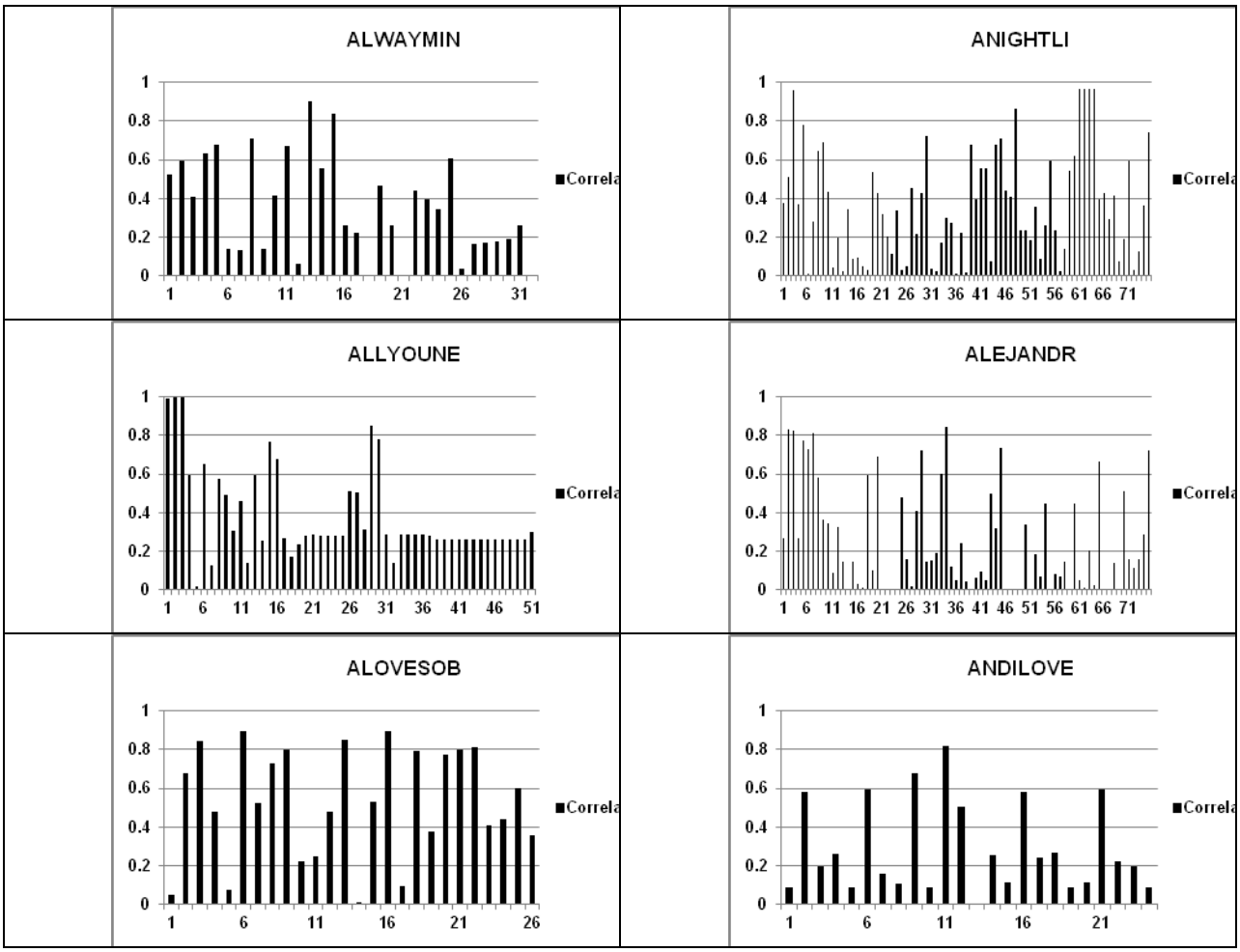


Figure 13: Evaluation using combined training data set of Twitter and 90 songs.

CHAPTER 6

DISCUSSION AND CONCLUSIONS

Through our experiments, I seek to determine the extent to which I can automatically determine the emotional category of each line in a song, for each of the six emotion dimensions.

A novel corpus of human-annotated emotions using Amazon Mechanical Turk at line level is introduced. Classifiers were built which could determine the emotion intensities in song lyrics using textual features. Another corpus of twitter data and the process of collecting it using Twitter API are explained in this thesis. Another system that helps in automated classification of emotions in song lyrics using an integrated data set of human-annotated emotions and twitter data is introduced.

Through the above-described experiments I see that correlation is high for most of the sentences in a song. There are variations in the results basing on different classification methods. For example, in the song “all you need,” the first three lines have “love love love.” These sentences have been classified as “joy” by human annotators. The classification method using bag of words, classify this sentence to be joy as well with correlation above 0.8. But these sentences have very low correlation (almost 0) in the naïve bayes classification.

I can see that the song “ASONGFRC” has high correlation values in all the sentences in the classification method-using bag of words and naïve bayes as well. While the song “ALLWAYMIN” has mixed correlation values some high and some low. If I closely compare these two songs “ASONGFRC” and “ALLWAYMIN,” the former song is high in “joy” while the latter has mixed annotations some in “sadness” and some in

“surprise” according to the human annotators. Similar is the case with the song “ALEJANDR,” where I have varied annotations from human annotators. Hence, the results don’t follow a pattern and have varied correlation values from the classification methods.

In the method, where Twitter data alone is used as the data set for naïve bayes classification, I see that the emotion-wise score for “joy” has the highest value when compared to the other five emotions. The tweets for “joy” were easy to collect using the Twitter API compared to the other five emotions. It was quite straightforward and easy to collect large number of tweets related to the key words like “happy,” “cheerful” or “joy.” It might be because these tweets were large in number compared to other five emotions. Hence, it reflects in the similar way to have a higher emotion-wise score for “joy.”

I have another system that combines the data set to be twitter data and human-annotated scores for the 90 songs. This classification system leads to a lesser impressive sentence correlation scores when compared to the other systems. However, the emotion-wise evaluation yields better results when using this integrated data set for the naïve bayes classification. The emotion-wise score for “surprise” has a higher score compared to any other emotion in any other method. I can interpret that the emotion-wise scores in this systems have significantly improved for emotions which had relatively low score in the method that uses twitter data alone while it didn’t impact much on “anger,” “disgust” etc.

Through experiments carried out using a dataset of 100 songs, I believe that emotions can be classified automatically. Automation of song lyrics in the 10 test songs

shows us good results in all the classification methods used. Classification methods with bag-of-words, the naïve bayes method and the classification method that uses twitter-data yields better results for the 10 test songs when compared to the classification that uses integrated data set. Twitter based method and the integrated data set yields better results in emotion-wise evaluation.

I can improve the systems if additional textual features and contexts are considered. In this thesis, the analysis of emotions in song lyrics is considered at sentence-level. The emotion flow of the songs can follow a pattern as the song flows. The flow of emotions might be consistent depending on whether it is a verse or a chorus. Also, it might be an interesting area of study to determine the emotions people are likely inclined to listen to. Considering audio features might enhance the systems as well.

REFERENCES

[1] Carlo Strapparava and Rada Mihalcea, Learning to Identify Emotions in Text, in Proceedings of the ACM Conference on Applied Computing ACM-SAC 2008, Fortaleza, Brazile, March 2008

[2] Y.-H. Yang, Y.-F. Su, Y.-C. Lin and H.-H. Chen, "Music emotion recognition: The role of individuality," ACM Int. Workshop on Human-centered Multimedia (ACM HCM), 2007.

[3] Y.-H. Yang, Y.-C. Lin, Y.-F Su, and H.-H. Chen, "A regression approach to music emotion recognition," IEEE Trans. Audio, Speech, and Language Processing, volume 16, number 2, pages 448-457, February 2008, (IEEE 2011 Young Author Best Paper Award)

[4] Yang Dan, Lee Won-Sook. Proc of the IEEE International Symposium on Multimedia. Washington, DC: 2009. Music Emotion Identification from Lyrics; pp. 624–629.

[5] C. Laurier, J. Grivolla and P. Herrera: “Multimodal Music Mood Classification Using Audio and Lyrics,” In Proceedings of the International Conference on Machine Learning and Applications, 2008

[6] Zhou, L.J., Lin, H.F., Liu, and W.F.: Enriching Music Information Retrieval Using Emotion Detection. In: SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011), Beijing, China, July 28 (2011).

[7] P. Ekman. An argument for basic emotions. Cognition and Emotion, 6:169–200, 1992

[8] Fabrice Colas, Pavel Brazdil: Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. IFIP AI 2006: 169-178

[9] An experimental study of Naïve Bayes Classifiers: I. Rish, “An empirical study of the naive Bayes classifier”, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001

[10] Michael I. Mandel, Dan Ellis: Song-Level Features and Support Vector Machines for Music Classification. ISMIR 2005: 594-599

[11] Tamsin Maxwell: Exploring the Music Genome: Lyric Clustering with Heterogeneous Features. page 37

[12] Rada Mihalcea, Carlo Strapparava: Lyrics, Music, and Emotions. EMNLP-CoNLL 2012: 590-599

[13] Mishne, G.: Experiments with Mood Classification in Blog Posts. In: 2005 Stylistic Analysis of Text for Information Access Conference (2005)

[14] S. Webb, and C. Pu. Study of Trend-Stuffing on Twitter through Text Classification. In Conference on Email and Anti-Spam (CEAS'10). Seattle, WA, July 2010.

[15] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A. and Turnbull, D. (2010). Music emotion recognition: a state of the art review. Proceedings of the 2010 International Society for Music Information Retrieval Conference, Utrecht, Netherlands: ISMIR

[16] Alexander Sorokin and David Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. In Computer Vision and Pattern Recognition Workshop

[17] Ipeirotis, Panagiotis. 2010. Analyzing the Amazon Mechanical Turk marketplace. XRDS.

[18] Chok, Nian Shong (2010) Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data. Master's Thesis, University of Pittsburgh.

[19] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of LREC 2010.