# An Examination of Cross-Cultural Similarities and Differences from Social Media Data with respect to Language Use

**Mohammad Fazleh Elahi, Paola Monachesi**
SFB/TR8 Spatial Cognition, University of Bremen, Germany
Utrecht University, Uil-OTS, Netherland
fazleh@informatik.uni-bremen.de, P.Monachesi@uu.nl

### Abstract

We present a methodology for analyzing cross-cultural similarities and differences using language as a medium, love as domain, social media as a data source and 'Terms' and 'Topics' as cultural features. We discuss the techniques necessary for the creation of the social data corpus from which emotion terms have been extracted using NLP techniques. Topics of love discussion were then extracted from the corpus by means of Latent Dirichlet Allocation (LDA). Finally, on the basis of these features, a cross-cultural comparison was carried out. For the purpose of cross-cultural analysis, the experimental focus was on comparing data from a culture from the East (India) with a culture from the West (United States of America). Similarities and differences between these cultures have been analyzed with respect to the usage of emotions, their intensities and the topics used during love discussion in social media.

**Keywords:** social media, cross-cultural analysis, Latent Dirichlet Allocation(LDA)

## 1. Introduction

Working in a cross-cultural environment and making decisions based on our own culturally specific treatment of an issue, idea or situation can result in miscommunication and ultimately, misguided outcomes. Culture analysis can provide us with insight into these concepts so that we may appropriately and effectively interact with people all over the world. To date, culture research is mainly conducted on the basis of a questionnaire- or survey-based approach, which is traditional, manual, time-consuming and carried out with respect to a limited number of respondents of each culture. Despite the importance of automatic analysis of culture, there is not much work that has been invested in this direction using social media data.

Language is the principal means whereby we achieve social interaction and it is bound up with culture in intimate and different ways. The words people use in communication reflect their expressions, ideas, beliefs and points of view. The research studies on 'emotion and culture' (Kitayama & Markus & Kurokawa, 2000) and 'cultural and love' (Gareis & Wilkins, 2009) show that culture has a great influence on love expression or discussion. Despite a significant amount of attention by academic researchers, as represented in various social media networks, to the best of our knowledge, no work has been invested in the field of culture analysis from social media through emotions analysis to date.

We present a methodology for analyzing cross-cultural similarities and differences using language as a medium, love as domain, social media as data source and 'Terms' and 'Topics' as cultural features. We discuss the techniques necessary for the creation of the social data corpus from which emotion terms have been extracted using NLP techniques. Topics of love discussion were then identified in the corpus by means of Latent Dirichlet Allocation (LDA) **(**Blei& Ng & Jordan, 2003**)**. Finally, on the basis of these features, a cross-cultural comparison was carried out. The experimental focus was on comparing data from a culture from the East (India) with a culture from the West (United States of America). Similarities and differences between these cultures was analyzed with respect to the usage of emotions, their intensities and the topics used during love discussion in social media.

This paper is organized as follows: section two focuses on previous research on cross-cultural analysis. Section three discusses methods and techniques of corpus creation from social media, culture feature extraction from the corpus and cross-culture analysis with respect to these features. Section four presents the hypothesis and the experimental setup for the cross-cultural analysis. Section five discusses cross-cultural similarities and differences identifiable with respect to topic and term analysis. Section Six contains some concluding remarks and future research directions.

## 2. State of the Art

Hofstede’s (1984, 2001) work on cultural dimensions has been regarded as the most promising paradigm in the field of cross-cultural research. From Hofstede’s research, we adopt the underlying idea of using cultural features in order to analyze a country's culture but we use a different set of cultural features and data medium, which is social media instead of questionnaires.

Kitayama et al. (2000) analyzed cross-cultural similarities and differences using emotions as cultural features. They concluded that the reported frequency of positive emotions was considerably higher than negative emotions for Americans. From this research, we adopt the underlying idea of using the positivity or negativity of an emotion and its strength as a feature for cross-culture analysis.

The goals of the research conducted by Gareis and Wilkins's (2009) and Horton et al. (2009) are similar to ours in that they analyze love expressions to understand cultural differences but they did not use social media as data source.

Nakasak et al. (2009) developed a visual interface from multilingual blogs with a topic keyword by analyzing cross-lingual/cross-cultural differences in concerns and opinions. The present study uses a similar approach but we analyze sub-topics of the main topic of research ('Love and Relationship') instead of random multiple topics of different kinds.

Nagarajan et al. (2009) analyzed similarities and differences between male and female profiles through term analysis of online dating systems. Paul and Girju (2005) presented an analysis of cultural differences from people's experiences from local and global perspectives. They developed an LDA (Latent Dirichlet Allocation) based probabilistic model and claimed that such a model provides a good framework for culture analysis. Unlike these approaches, we aimed to take culture analysis to a new dimension by employing a dictionary of emotions constructed from widely used emotion corpora and different types of sub-topics of the main topic of research ('Love and Relationship').

## 3. Methodology

In this section, a methodology for culture analysis using 'love and relationship' as a domain, social media as data source and language as a medium is outlined. The overall architecture, which is shown in figure 1, comprises three modules.

More specifically, social media data such as forum discussions and blogs are crawled and a corpus is created that can be further analyzed with respect to culture i.e *Corpus Creator* module. Furthermore, cultural features are extracted. In our case, the most frequent topics in our corpus of love blogs were investigated. We used Latent Dirichlet Allocation (LDA), an unsupervised machine learning technique for extracting all latent topics and from these, we identified the most frequent ones by using a ranking mechanism i.e. *Feature Extractor* module. In addition to topics, we take emotion terms into account and we consider different levels of emotions and their strengths. To this end, an emotion lexicon was created on the basis of three widely used emotion corpora, such as SentiWordNet3, [1] WordNetAffect1.1 [2] and SemEval 2007.[3] Finally, the features extracted were classified into three groups (*All*, *Shared* and *Unique*) for the analysis of similarities and differences across cultures i.e. *Culture Feature Analyzer* module.

### 3.1 Corpus Preparation from Social Media

In order to carry out culture analysis on the basis of social media data, a corpus is necessary. It should be (i) large in volume, (ii) domain specific (iii) and with geo-Location Information. Unfortunately, we were not able to find an available social media corpus in the domain of our study given that geo-location information is often missing. It is for this reason that we have developed a corpus building tool i.e *Corpus Creator* module. Figure 2 presents the corpus' building tool that consists of five components. More specifically, (i) The *Post URL Crawler* takes a URL of a social media site and returns the URLs of all the posts or messages. (ii) The component *Post Country Extractor* takes each of the URLs of user posts or messages and extracts user provided geo-location information. User provided location is then mapped to a country. If the country of origin is known then this step is ignored. (iii) The component *Country-URL Mapper* takes the country information of a post with its URL and makes a tuple of the pair (country, post URL) for all posts. This information is the input of the *Culture Divider* module. (iv) The *Culture Divider* groups this tuple with respect to the culture groups necessary for our analysis and it feeds to the *Post Content Extractor*. (v) The *Post Content Extractor* explores each post URL (or link) of a culture and extracts 'post content'. Finally, the country information and 'post content' are stored in files in XML format.

The *Post Country Extractor* is the most important component since geo-location information is essential in order to carry out analysis of culture. Three mapping wordlists have been created in order to assign a user provided location to its country. The first list contains mapping information from major locations to their corresponding countries. For preparing this list, we use a crawler that travels Wikipedia pages containing state-country, city-country, capital-country information and extracts the mapping list using regular expression. A second list is built from the MaxMind[4] location dataset that contains the mapping information from a location to country. A third list is created on the basis of the manual analysis of the noise in user provided information: it maps incomplete, misspelled, abbreviated or compressed forms of some major cities and countries to their corresponding country.

The *Corpus Creator* has been employed to assemble data from a Love Forum called '1 million Love Message'[5] and from the blogcatalog.com. 9406 posts from 253 Indian blogs and 10123 posts from 7653 American blogs were utilized for our experiments. The reason for using different number of posts instead of normalizing was to keep the number of terms extracted from both cultures approximately equal.

### 3.2 Culture Feature Extraction

In our research, we analyze cross-cultural similarities and differences by employing language used in blogs and forum posts. In particular, we analyze (i) topics and (ii) terms attested in our corpus.

#### 3.2.1. Topics Extraction from Social Media Corpora

A post is a collection of words and a topic can be defined

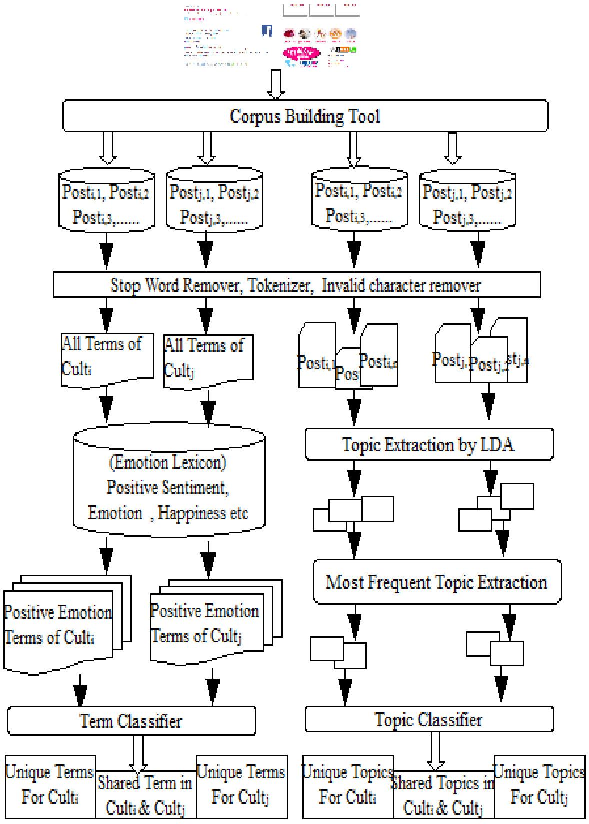
[1] http://sentiwordnet.isti.cnr.it/

[2] http://wndomains.fbk.eu/wnaffect.html

[3] http://nlp.cs.swarthmore.edu/semeval/

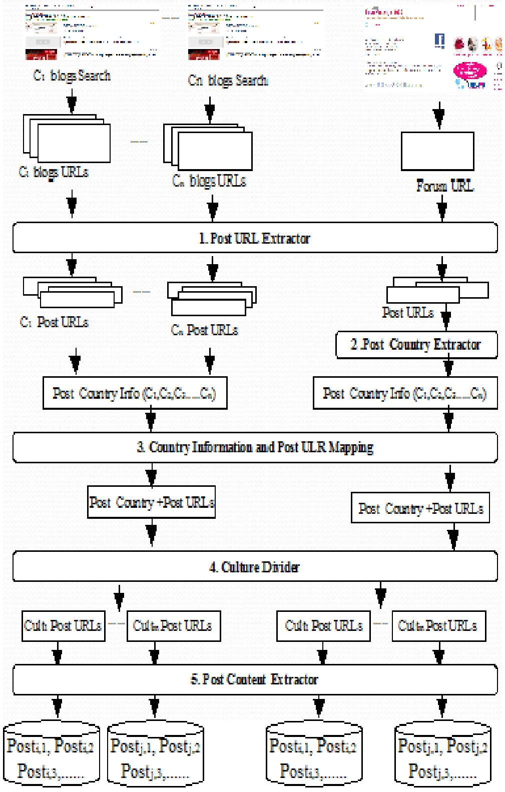
[4] http://www.maxmind.com/app/worldcities

[5] http://www.1millionlovemessages.com/



Corpus Building Tool
Posti,1, Posti,2 Posti,3,......
Postj,1, Postj,2 Postj,3,......
Stop Word Remover, Tokenizer, Invalid character remover
All Terms of Culti
All Terms of Cultj
Posti,1
Posti,4
Postj,1
Postj,4
Topic Extraction by LDA
(Emotion Lexicon) Positive Sentiment, Emotion , Happiness etc
Most Frequent Topic Extraction
Positive Emotion Terms of Culti
Positive Emotion Terms of Cultj
Term Classifier
Topic Classifier
Unique Terms For Culti
Shared Term in Culti & Cultj
Unique Terms For Cultj
Unique Topics For Culti
Shared Topics in Culti & Cultj
Unique Topics For Cultj

Figure 1: The overall architecture of culture analysis from social media.



C1 blogs Search
Cn blogs Search
C1 blogs URLs
Cn blogs URLs
Forum URL
1. Post URL Extractor
C1 Post URLs
Cn Post URLs
Post URLs
2. Post Country Extractor
Post Country Info (C1,C2,C3.....Cn)
3. Country Information and Post ULR Mapping
Post Country +Post URLs
4. Culture Divider
Culti Post URLs
Cultn Post URLs
5. Post Content Extractor
Posti,1, Posti,2 Posti,3,......
Postj,1, Postj,2 Postj,3,......

Figure 2: Corpus creation from Forum and Blogs.

as a collection of words taken from the post. The fundamental idea of topic modeling (Griffiths & Steyvers, 2002) is that a document consists of several topics, similarly in a blogpost, therefore the single topic approach behind the Näive Bayes model (Mitchell, 1997) is not adopted in this study. In pLSI, (Hofmann, 1999) each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads to a problem as the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting (Blei & Ng & Jordan 2003). In culture analysis from social media, the posts or messages of different cultures are used as corpus, so the corpus can be very large in size and therefore, pLSI is not suitable. Therefore, for this study, the Latent Dirichlet Allocation (LDA) model, in which all the above limitations are addressed is used. A Gibbs sampling approach is utilized to learn the LDA parameters, which is computationally efficient and thus important for large data set.

Our goal is to take a corpus of a given culture, which consists of blogposts for $Cult_i$ and to return the most used topics in that culture. i.e $Topic_i$= ($T_{i,1}$, $T_{i,2}$,....,$T_{i,L}$) where L is the number of most used topics in a culture. Our topic extraction task consists of the following two steps. Firstly, for the corpus of a culture, we extract text from each blogpost and then tokenize each word and remove stopwords and invalid characters. This corpus is sent to the next step for latent topic extraction. After that, we run LDA over all the blogposts of a culture in order to discover topics automatically. Our study uses the Java-based package Mallet[6], as it uses Gibbs sampling for parameter estimation and computational efficiency. We run the tool for 200 topics with 1000 sampling iterations. The number of topics depends on what we are looking for in the model. Therefore, the number of topics used for topic analysis depends to some degree on the size of the collection. The Mallet LDA tool takes the corpus of a culture $Cult_i$ as input and produces two outputs: (i) Z (Z=200) latent topics and the top k (k=50) top words for each topic. These topics are named manually by looking at the top 50 words of each latent topic. (ii) the proportions of each topic of these Z (Z=200) topics in a blog post.

Secondly, the output of Mallet is post-processed in order to find the most used topics of a culture in love discussion. Therefore, we sort the proportions of all Z topics of a post in descending order and delete the topic proportions that are negligible with respect to other topics. For deleting unimportant topics for a post, we delete the topic proportion that is less than a threshold (0.1). After that, we make a tuple <U,T> for each post where U is the URL of the post and T={($t_1$,$p_1$)), ($t_2$,$p_2$)),....... ($t_t$,$p_t$)} I.e $t_1$, ..,$t_n$ are topics and $p_1$,......, $p_t$ are topic proportions that remain after removing the less important topics of that post. After that, a tuple is created which contains the topic and the number

[6] http://mallet.cs.umass.edu/

of posts $<T,N>=(t_1,n_1),(t_2,n_2),(t_3,n_3).......(t_t,p_t)$ where T is the set of topics and N is the set of the number of times the topic is found in the posts of a culture. The topics are then ranked in descending order.

#### 3.2.2. Term Extraction from Social Media Corpora

Emotion is a small subset of sentiment. For example, 'life', 'power', 'love', 'death', 'sorry' are sentiment words but, among them, only 'love' and 'sorry' are emotion words. Emotions are widely used in sentiment analysis research for analyzing sentiments of different groups of people (Pang & Lee, 2008).

In order to carry out an analysis of different types of emotions, we built an Emotion Lexicon on the basis of three widely used emotion corpora: SentiWordNet3.0 (Baccianella & Esuli & Sebastiani, 2010), WordNetAffect1.1 (Strapparava & Valitutti, 2004), and SemEval 2007 (Strapparava & Mihalcea, 2007). We used the SentiWordNet3.0 lexicon to classify the terms according to their polarity (positive and negative) and the intensity (strong, medium and weak) by relying on the terms' scores associated in the lexicon. For extracting terms related to positive and negative emotion, we follow the path of emotion hierarchy of WordNetAffect1.1 and assign its root value as its category. For example, 'carefreeness', 'cheerfulness' and ‘joy' are assigned to Positive Emotion. For the purpose of the analysis of different types of basic emotions in love discussion, a list of words related to six basic emotions, namely, Anger, Fear, Happiness, Sadness and Surprise, was also constructed from SemEval 2007.

In order to extract emotion terms from social media corpora, first, we take all the posts of our corpus as input and remove stopwords and invalid characters and punctuation. Secondly, we consider the corpus of a culture as a document, so a document contains all the terms of a culture. After that, we calculate the number of occurrences of a term. Thirdly, text samples of a culture are analyzed on a word-by-word basis, comparing each word to those in emotion categories in the Emotion Lexicon. If the terms match any of these emotion groups, then they are recorded. Each term can match with one or more sentiment and emotion categories. For example, the term *joy* is part of three emotion categories: positive sentiment, positive emotion and happiness emotion. In this way, a tuple is prepared which consists of the set of emotion terms found in a culture. Each set for each emotion group consists of the terms and the number of occurrences of these terms in a culture.

### 3.3 Cross-Cultural Analysis through Language

For the purpose of cross-cultural analysis, we classified the cultural features into three groups: (i) All the terms or topics found in a culture are called *All Terms* or *All Topics*. (ii) Terms or topics commonly found in all cultures are called *Shared Terms* or *Shared Topics*. Emotion terms like 'love', happiness' or 'sadness' are usually commonly found in love discussions of all cultures so they are *Shared Terms*. The topics 'Marriage', 'Relationship', 'Emotion' are usually commonly found in love discussion of all culture so they are *Shared Topics* and (iii) the Terms or Topics that are only found in one culture are called *Unique Terms or Unique Topics*. For example, the *'Hindi Words for Love'* is only found in love discussions of Indian Culture; therefore, the topic is a *Unique Topic* for India.

## 4. Experimental Setup

The methodology has been tested on the basis of several experiments aimed at comparing data from a culture from the East, which is India and a culture from the West, which is United States.

It was hypothesized that in discussing about love (i) *there would be clear differences in most frequently used topics in both culture.* (ii) *Indians and Americans are likely to discuss topics that will be related to the particular traditions and recent events in their culture.* (iii) *the use of positive terms would be higher than the negative ones in both cultures and* (iv) *Indians would be more emotional than Americans.*

In order to test these hypotheses, first, we analyzed the most frequent topics from *All topics* found in a culture and compared them with the other culture. Tables 1 and 2 present the top ten topics found in Indian and American blogposts, respectively. For each topic, top frequent terms are also shown. For example, the most highly frequent topic found in Indian culture is 'Hindi Love Emotion Terms' and the top frequent terms for this topic are *'maiy ek kabhi mere par pyar hu yaad har....'*.

Second, we analyzed cross-cultural similarities and differences with respect to the extracted *Unique Topics*. In order to identify the *Unique Topics*, we created tuples $Cult_i<Topic_i,Term_i>$ where $Topic_i$ is a *Unique Topic* that is found in $Cult_i$ but not in other culture and $Term_i$ is a set of terms of that topic. Table 3 and table 4 exhibit *Unique Topics* found in both India and the USA.

Third, we considered the topics that are commonly found in love discussions of all cultures. On the basis of the terms used in these topics, a cross-cultural comparison is carried out. For the purpose of the analysis, *Shared Topics* from all cultures were extracted and then *Shared Terms* were deleted from them so that *Shared Topics* finally consisted of only *Unique Terms*. Table 5 presents two *Shared Topics* found in both cultures and the *Unique Terms* of those topics in each culture. For example, the topic 'Same Sex Issue' is a *Shared Topic* between India and America. So, SameSex <India (gay, delhi, supreme….), USA (gay, president, military….)>. As the term 'gay' is a *Shared Term* it was deleted. After that only *Unique Terms* i.e SameSex <India (delhi, supreme,....), USA (president, military....)> remained.

Fourth, we analyzed the rate of use of terms of an emotion category in a culture. For example, we calculated the use of happiness emotion terms and their proportion over all terms of a culture. In Table 6, the percentages of usage of different emotion categories in India and USA are shown.

Fifth, we investigated cultural similarities and differences considering *Shared* Sentiment and Emotion Terms.

| Topic | Most frequent words in the topic |
|---|---|
| Hindi Love Terms | maiy kabhi pyar hu yaad pyaar hota bahut raat hoti ….. |
| Hindi Emotion Terms | nahi ko ne meri hum bhi tera apne pal bewafa mujhe …. |
| Wedding Ceremony | wedding bride ceremony groom flower cakes couples clothes responsibility music .. |
| Relationship Terms | relationship partner trust intimate share mutual strong commitment .. |
| Dating Terms | dating meet people tips confidence dates time personality advice ….. |
| Love Emotion Terms | love feeling hurts ocean hide met falling attention faith heart...... |
| Online Dating Terms | online sites services internet share categories posts photos faith seeking moment........... |
| Negative Emotion | love pain cry infatuation trust wait patient lie quiet public accepting …. |
| Pregnancy Terms | infertility fertility female pregnancy health risk disease baby child........ |
| Sexual Terms | sex sexual partner excitement satisfying desire emotional play marital …........... |

Table 1: Top ten topics found in love discussion in India.

| India Topics | Terms of the Topics |
|---|---|
| Topic 128 | ki se bhi na ka koi nahi ek mein par ye meri yaad ke tha kuch tum pyaar ne mere jo baat hai mujhe…. |
| Topic 187 | hai ke maiy kabhi pyar har hu mai ho gam mere jab bahut ka mobile hoti experienced… |
| Topic 110 | india sex government society indian section law ruling decision religious sodomy imprisonment …. |
| Topic 59 | female fertility weight baby wheat strangers child absence prevents worst websites realistic grounds…. |

Table 3: Unique Topics in India.

| Topic | Unique Terms of the Topics in India | Unique Terms of the Topics in USA |
|---|---|---|
| Wedding | Decoration,Responsibility, Duties, Rituals, Wishes, Traditions, Memorable, Cloth, Happiness... | Photographer, Law, Drinks, Bridesmaid,Gown, Skirt, Candle, Card, Suit, Snowflake …. |
| SameSex Issue | Delhi, supreme, article, mental, judgment, eye, ground, march view, mind, movies, nice, realize, sound straight.. | President, discrimination, orientation, military, community, policy, florida, defense, officials, lgbt, human, America, mexico, Washington, California, activists.... |

Table 5: Unique Terms of Shared Topics

| Emotion Group | India(%) | USA(%) |
|---|---|---|
| Strong Sentiment | 1.61 | 1.83 |
| Medium Sentiment | 30.69 | 26.57 |
| Weak Sentiment | 73.01 | 76.93 |

Table 7: The usage of Shared sentiment terms with respect to intensity in India and USA

| Topic | Most frequent words in the topic |
|---|---|
| Love Emotion Terms | love true falling loves soul feeling connected convinced honest hurt mental truth... |
| Dating Terms | call dating women messages mails posts site profile meet online match singles daters.. .. |
| Love Partner Terms | boyfriend break relationship girlfriend page jealous eventually completely fast.... |
| Wedding Invitation | invitations wedding print invites bridal cards printing custom stationery designer formal .. |
| Marriage Law | sex couples marriage gay law bill california lesbian marriages congress court ... |
| Wedding Venue | wedding venue choose theme guests suit planner garden restaurants site …. |
| Same Sex Issue | marriage gay legal unions support legislation discrimination homosexuality partner hate... |
| Wedding Flowers | flowers wedding bouquets florist floral bridesmaids colors cute drinks......... |
| Relationship Terms | women relationship commitment compatibility commit potential ........... |
| Love Positive Emotion Terms | love god fear amazing faith soul intentions afraid grow laughing dreaming ….. |

Table 2: Top ten topics found in love discussion in USA.

| USA Topics | Terms of the Topics |
|---|---|
| Topic 60 | partners department sex domestic clinton foreign secretary office service policy health federal personnel house government act affairs diplomats .. |
| Topic 29 | flowers bouquets silk arrangements colors floral beautiful color decoration roses orange ceremony .. |
| Topic 190 | romantic candles ideas light romance winter snow wine images natural inspiration stood static answer .. |
| Topic 177 | immigration country act homosexuality born sponsor status local building marrying victim.. |

Table 4: Unique Topics in USA.

| Emotion Group | % (India) | % (USA ) |
|---|---|---|
| Positive Sentiment | 30.26 | 22.06 |
| Negative Sentiment | 24.17 | 19.27 |
| Positive Direct Emotion | 2.13 | 0.64 |
| Negative Direct Emotion | 0.27 | 0.15 |
| Happiness Emotion | 3.13 | 1.41 |
| Sadness Emotion | 0.45 | 0.22 |
| Anger Emotion | 0.26 | 0.17 |
| Fear Emotion | 0.14 | 0.11 |
| Surprise Emotion | 0.19 | 0.13 |
| Disgust Emotion | 0.03 | 0.03 |

Table 6: The usage of all emotion terms in India and USA

Knowing that these terms are commonly used by cultures, focus was placed on their intensity after which a cross-cultural comparison was carried out. We calculated the usage of *Shared* Strong/Medium/Light Sentiment terms over all *Shared* sentiment terms in a culture. Table 7 presents the use of Strong/Medium/Light Shared Sentiment Terms in India and the USA respectively.

## 5. Results and Discussion

By looking at the most frequently used topics in table 1 and table 2, it is possible to compare which types of topics are popular in each culture. For example, it was found that four topics out of ten topmost topics of Indian blogposts are Emotion topics. Along with these differences, there are also similarities in love discussion between them. Some topics like (Dating, relationship etc) were found to be equally important in both cultures.

From the Table 3, some of the *Unique Topics* identifiable in Indian blogposts are Hindi words written in English

(Topic128, Topic187). It is highly unlikely that such topics will be found in American blogs. Other *Unique Topics* found from Indian blogs were Topic110 and Topic59 related to the Indian sexual law issue and pregnancy issue, respectively. On the other hand, *Unique Topics* in the USA are Topic60 which is related to sex and political discussion of US government, Topic177 is related to the homosexual immigration issue. All these

*Unique Topics* are culture-specific dealing with issues which only exist in that culture. Therefore *Unique Topics* are culturally-significant reflecting the characteristics of that culture.

From the Table 5, some of the *Unique Terms* (*delhi*, *supreme*, *judgment*, *march*, *mental* and so on) of the *Shared Topic* (Same Sex Issue for India) are related to the discussion on the recent declaration of Delhi (India) High Court which 'decriminalized homosexuality and march of LGBT activists in the support of this declaration'(BBC News)[7]. On the other hand, some of the *Unique Terms* (*president, military, community, policy, florida, defense, official, lgbt, American, human, rights Washington, California, activists* and so on.) of the *Shared Topic* SameSex Issue for USA are the name of the locations of United States of America probably because most of the discussion related to the Same Sex Issue in the USA are related to discrimination against or acceptance of such relationships in different provinces. The terms *president, military, community, policy, defense, official* and so on are related to the controversy of the policy of same sex relationship in USA military or between defense officials(Don't ask, don't tell)[8]. Therefore, Indians and Americans are likely to address topics in love discussion that will be related to the particular traditions and recent events in their cultures.

From the Table 6, in both cultures, the use of different types of positive terms (sentiment and emotion) is higher than the negative ones. Also for all emotion categories, the percentage is higher in India than in the USA. Results show similar findings to psychological research (Ekman, 1972) on culture analysis using emotions as features while other contrasts with their outcome (Kitayama & Markus & Kurokawa, 2000). On the other hand, Table 7 shows that both Indians and Americans use Weak Sentiment Terms more than Medium Sentiment Terms and Strong Sentiment Terms in love discussion.

All these results show that this kind of analysis could support standard, traditional methods of research and provide new interesting insights of culture.

## 6. Conclusion and Future Work

The approach presented in this paper has several advantages for culture analysis. There are clear benefits in using social media for culture analysis since there is availability of a huge volume of data in several domains from all geographical distributions. Cultural features can be automatically extracted and hypotheses can be tested using minimum human intervention. The methodology presented can be applied to any culture in any geo-location and it can be used for any domain. There are several areas of application for this research. Culture analysis with respect to emotions using social media can add value to the research in the behavioral sciences and emotion research.

As the language is the principle means whereby we achieve social interaction and it is bound up with culture in intimate and different ways, so people can express views, ideas and emotional expression more clearly in their own languages. Our methodology of culture analysis was employed on English written text of love for both India and USA. Therefore, cross-culture analysis with respect to emotions can be conducted on multilingual social media to get better insights. In our research, we used a knowledge-based approach for emotion extraction. Along with word-level analysis, emotion analysis can be extended to a corpus-based approach by using a collection of blogposts annotated with basic emotions (Strapparava & Mihalcea, 2008). Emotions have certain features such as expressive utterances, exclamatory utterances, expressive sentences, and exclamations. For example, the following sentence may not contain emotion terms but it is an emotionally-expressive sentence: “How very curious!” (Shaw, 1941).

Culture analysis with respect to emotions can be also applied to different fields of a culture, for example, emotion analysis of male-female or teenager-adult interaction in a culture. Computational models of culture are valuable for designing complex open systems such as serious games, social simulations, personal assistants, collective intelligence and social network software. On the basis of cross-cultural analysis from social media, a computational model of culture can be developed using semantic web technology.

## 7. References

Baccianella, S., Esuli, A., Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, MT, 2010, pp. 2200-2204.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent to Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), Nebraska Symposium on Motivation 1971, Vol. 19, pp. 207-283.

Gareis, E. Wilkins, R. (2009). Emotion Expression and the Locution "I Love You": A Cross-Cultural Study". Paper presented at the annual meeting of the International Communication Association, Sheraton New York, New York City.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24 th Annual Conference of the Cognitive Science*

[7] http://news.bbc.co.uk/2/hi/8129836.stm

[8] http://en.wikipedia.org/wiki/Don't_ask,_don't_tell

*Society.*

Hofmann, T. (1999). Probabilistic latent semantic indexing. In SIGIR ’99: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA. ACM.

Hofstede, G., Bond, M. (1984). Hofstede's culture dimensions: An independent validation using Rokeach's value survey. *Journal of Cross-Cultural Psychology*, 15, 417-433.

Hofstede, G. (2001). Culture’s consequences: Comparing values, behaviors, institutions, and organizations across nations. Thousand Oaks, CA: Sage.

Horton, B. W., Kline, S., Zhang, S. (2009). "How We Think, Feel and Express Love: A Cross-Cultural Comparison Between American and East Asian Cultures". Paper presented at the annual meeting of the International Communication Association, Sheraton New York.

Kitayama, S., Markus, H.R, Kurokawa, M. (2000). Culture, Emotion, and Well-being: Good Feelings in Japan and the United States. Cognition & Emotion, 1464-0600, Volume 14, Issue 1, 2000, Pages 93 – 124.

Mitchell, T. (1997). Machine Learning. McGraw-Hill, Boston.

Nagarajan, M., Hearst, M. (2009). An Examination of Language Use in Online Dating Profiles. *Proceedings of the Third International ICWSM Conference (2009).*

Nakasak, H., Kawaba, M., Yamazaki, S. (2009). Visualizing Cross-Lingual/Cross-Cultural Differences in Concerns in Multilingual Blogs. *Proceedings of the Third International ICWSM Conference*.

Paul, M., Girju, R. (2009). Cross-cultural analysis of blogs and forums with mixed-collection topic models. In EMNLP ’09: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417, Morristown, NJ, USA, 2009. Association for Computational Linguistic.

Pang, B., Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.

Shaw, G. B. (1941). Pygmalion. Harmondsworth: Penguin.

Strapparava, C., Mihalcea, R. (2008). Learning to Identify Emotions in Text. *In SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1556 - 1560.

Strapparava, C ., Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text, *Fourth International Workshop on Semantic Evaluations (SemEval-2007),* ACL, 2007, pp. 70-74.

Strapparava, C., Valitutti A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proc. of 4th International Conference on Language Resources and Evaluation (LREC 2004),* Lisbon, May 2004, pp. 1083-1086.