Psyche Extraction Accuracy of English Texts of Non Native Speakers in Pakistani Universities

Summaira Sarfraz
Department of Sciences and Humanities
National University of Computer and Emerging Sciences (FAST-NUCES)
Lahore, Pakistan

Ahsan Nabi Khan
Department of Computer Science
National University of Computer and Emerging Sciences (FAST-NUCES)
Lahore, Pakistan

Abstract

The paper, based on Linguistics and Text Mining integrated research, aims to investigate the problem of determining the accuracy of the overall psyche of a text. It examines not only how a particular text can be labeled by a dominant mood, but also handles the complexities of multiple moods in the same context, ranking by scores of extracted psyche categories. The psyche scores are based on occurrences of related keywords and their intensity weightages. A text mining tool 'Psyche Map' has been developed for the extraction of words associated with moods from the text and the determination of the text psyche based on the calculation of the weightages attached to these words associated with moods. The tool is tuned to tag local psyche on account of its basis in determination of moods, their synonyms and their respective weightages, which was undertaken locally from non-native English language teachers who teach at undergraduate level. The instruments of the study are the English essays written by the non-native undergraduate university students of English language course and the psyche of their texts have been accurately determined by the tool on the basis of moods interpretation.

Keywords*-Linguistics; Text Mining; Accuracy; Overall Psyche; Moods Interpretation*

1. INTRODUCTION

The purpose of this study is the interpretation of the words associated with moods based on their intensity, the contribution of other neutral words and multiple emotions within the same text to form a particular psyche. The study, based on an integrated Linguistics and text mining research, aims to examine the psyche of English texts written by non-native Pakistani undergraduate university students of English language. We take sentiments and emotions as moods of a writer. Knowing a mood means knowing a psyche of a person when he/she is in the process of writing. The mood is commanding a writer's writing and thought process and is reflected in the text. Thought process is as varying as mood of a person, but there are patterns when such mood variations repeat themselves.

The study further determines the accuracy by looking at a writer's own claim of the mood expressed in the text. This paves a way to assess the prediction of the likely mood.

2. Literature Review

The extensive research on automatic text analysis for moods and sentiments to detect the psyche of a text, based on Sentiment Analyzer, Affective Text Analysis Approach, Opinion Extraction or Recommender Systems typically extract the overall sentiment revealed in a document, either positive or negative, or somewhere in between (Jeonghee 2003) .The challenges faced by sentiment analysis are that though the overall opinion about the subject is useful but sentiments about the individual aspects of subject are also to be taken into account. In reality, for example, though one can be generally happy about his car, he might be dissatisfied by the engine noise (Jeonghee 2003).

One of the objectives of affective computing is to recognize human emotions. Research has shown that affect recognition not only provides understanding of a cognitive processes of human mind but also assist us in making computers that better react to human input. Existing research on affect recognition has mainly looked at sources of facial expressions, vocal intonation and physiological signals that vary with affective states. However, although language is an important way to convey emotions, either in oral forms such as public speeches or in written forms such as prose or poetry, little work has been conducted on detecting emotions based on verbal expressions (Leshed 2005).

Researchers over the years have proposed that there exist between two and twenty basic emotional categories but there is a difference of opinion on the final count due to the synonymity of moods. There is still no final word on the hierarchy cut for weighing the synonymous moods for classification. Although research has proved that emotional state of writer's mind can be predicted to some degree but the question of accuracy still remains; when a writers selects a mood, his purpose might not be to identify his/her emotional state and correlate it with writing.

Moods are very important part of what makes people read and browse different blogs. This growing importance of moods detecting from the huge corpus available online as has encouraged researchers to develop programs/tools tailored to moods identification. Latest development of MoodViews online aims to develop novel methods for searching, discovering and retrieving blogs. It consists of three tools; Moodgrapher, Moodteller, and Moodsignal, for mood tracking of text available online by LiveJournal. Though useful, its function is limited as the number of moods that the tools track is not sufficient to cover the overall psyche of the text. Also the tracking is mostly done on the basis of the occurrences of a particular word in a given time.

3. METHODOLOGY

The methodology of the research was based on four tasks:

1. Determination of mood categories, their synonyms and related words
2. Determination of intensity weights of moods and their synonyms
3. Sample essays written by non native university students of English language.
4. Accuracy check by the text mining tool "PsycheMap"

To determine the overall psyche of a text, the understanding of the intensity of the dominant mood was essential. Since the study is focused on analyzing the local psyche, each legend mood and its synonyms were ranked from lowest to highest based on their intensity by 40 non-native English language teachers. Each word's intensity was assigned a weight; lowest (1), low (2), neutral (3), high (4), highest (5).

Fifty sample essays were collected from the non- native university students of English language. These students were asked to write expressive essays and were further asked to verify their dominant mood(s) in their essays. The essays were then processed in PsycheMap for determination of the overall psyche of these essays.

4. IMPLEMENTATION

'PsycheMap' is a text mining tool which has been specifically developed for the study by computer programmers for tagging and calculating the overall psyche of a document. In PsycheMap, occurrences of any of the psyche words, their recorded synonyms or variations are calculated and incorporated in a total score of the document psyche.

4.1 Functional Specifications

Identify and record tokens of the English writing text. We use the whitespace and punctuation marks as delimiters. The token contains information of the lexeme (instance of the word), the category where it belongs, the weight and the count. It constructs a queue of the tokens that contain predefined psyche words, synonyms and word variants.

1) makeMap() reads from the predefined psyche words file and constructs the queue, that is, it creates a map (queue) by reading psyche words from file.

2) mapFile() reads the input file that is any English text. It tokenizes the text and counts occurrences of psyche words and save the new count in the queue.

3) calcWeight() calculates the score of overall psyche of documents (any English text) based on the formula below where W is the predefined weight and C is the count of the word form in the whole document:

4) $$Score = \sum_{n} W_i \times C_i \square$$

4.2 Input Files

1) "psychewords.txt": This is the file having the list of psyche words, synonyms and variations of word forms and associated weights as assigned by 40 non native English language teachers. It is saved as a tab-delimited file, to be opened as a spreadsheet.

2) "input.txt": This file contained any English writing text to be tagged.

4.3 Output Files

1) "psycheWeightOutput.txt": This file contained the list of psyche words categories (legend) and their respective total score from the document as calculated by the function calcWeight().

2) "checkedWords.txt": This file contained the list of all psyche words synonyms and word forms, associated with the legend, the count and the weight. This file is used to help the tester manually calculate the scores from the respective weights and counts.

3) "dataClip.txt": This file creates a dataset for analysis in Data Mining tools. It creates a comma-separated file containing numbers representing scores in the order of psyche word categories as occurred in "psychewords.txt".

5. RESULTS

Out of 50 essays, 18 showed accurate results as per the writers' claims of their dominant moods as shown in **Table 1.** The text mining tool determined the overall psyche of the essays exactly as expressed by the writers based on the occurrences and the calculated intensity weightages of the expressed moods.

Twenty essays showed results close to accuracy as expressed by writers. Quite similar, here, means that the expressed moods are either synonyms of the results (predicted moods) or fall under that result category or they give the same meaning that predicted moods provide.

Seven essays showed inaccurate results i.e. the results were entirely different from what the writers expressed as their dominant moods.

Five essays results could not be expressed as the writers of these essays did not specify their dominant moods in the essay. However, the text mining tool provided the results as per its procedure.

TABLE I. PSYCHEMAP RESULTS OF 50 NON-NATIVE STUDENT S'ESSAY RESPONSES WITH EXPRESSED AND PREDICTED MOODS

Accurate Results			
Essays	**Specified Mood(s) by the Writers**	**Predicted Moods with Maximum Frequency**	**Highest Frequency with Intensity**
1	**Confident**	**Confident**	**15**
2	**Satisfied**	**Satisfied**	**7**
3	**Hopeful**	**Hopeful**	**9**
4	**Depressed**	**Depressed**	**15**
5	**Powerful**	**Powerful**	**23**
6	**Confident**	**Confident**	**25**
7	**Anxious**	**Anxious**	**25**
8	**Anxious**	**Anxious**	**27**
9	**Sad**	**Sad**	**10**
10	**Glad**	**Glad**	**15**
11	**Bored**	**Bored**	**16**
12	**Proud**	**Proud**	**12**
13	**Scared**	**Scared**	**14**
14	**Confidence**	**Confident**	**15**
15	**Confident**	**Confident**	**17**
16	**Confident**	**Confident**	**20**
17	**Confident**	**Confident**	**18**
18	**Satisfied**	**Satisfied**	**17**
Results Close to Accuracy			
1	**Depression, Happy**	**Glad**	**13**
2	**Free, Calm, Comfortable, Hopeful**	**Friendly**	**16**
3	**Excited, Eager, Pressured, Uncomfortable**	**Anxious**	**23**
4	**Confident, Overwhelmed, Proud**	**Great**	**8**
5	**Charmed, Determinant, Hopeful**	**Friendly**	**16**
6	**Cautious, Confident**	**Comfortable**	**21**
7	**Cheerful**	**Glad**	**7**
8	**Confident, Overwhelmed, Proud**	**Satisfied**	**11**
9	**Confident, Satisfied**	**Comfortable**	**11**
10	**Depressed yet Relaxing and Exciting**	**Anxious**	**13**
11	**Concerned and Energetic**	**Competitive**	**9**
12	**Energetic, Tired, Uncomfortable**	**Jumpy**	**19**
13	**Charmed, Determined, Hopeful**	**Comfortable**	**14**
14	**Confused, Happy, Helpful**	**Satisfied**	**15**
15	**Bored, Indifferent, Uncomfortable, Satisfied and Charmed**	**Friendly**	**16**
16	**Satisfied and Happy**	**Friendly**	**20**
17	**Bold, Brave, Capable, Confident**	**Friendly**	**12**
18	**Joy, Confidence and Self-esteem**	**Confident**	**15**
19	**Glad, Happy and Satisfied**	**Friendly**	**16**
20	**Confident**	**Comfortable**	**14**
Inaccurate Results			
1	**Depressed, Hopeful**	**competitive**	**11**
2	**concern, double fullness, scared**	**confident**	**26**
3	**proud, burdened and pressurized**	**helpful**	**12**
4	**confident and scared**	**friendly**	**12**
5	**Good, Comfortable, Confident**	**jumpy**	**32**

6	**anxious, uneasy, and uncomfortable**	**friendly**	**20**
7	**Glad, cooperative**	**anxious**	**22**
Results without Writer's Specification of their Moods			
1	**<not expressed>**	**jumpy**	**19**
2	**<not expressed>**	**confident**	**9**
3	**<not expressed>**	**confident**	**18**
4	**<not expressed>**	**confident**	**20**
5	**<not expressed>**	**anxious**	**23**

6. DISCUSSION

The accurate results and results close to accuracy indicate that the writers have been able to use the relevant moods related words with right intensity. Even the occurrences of these words were quite in the line of the process of the text mining tool which has enabled the tool to determine the overall psyche accordingly. In 5 out of 50 essays, the writers did not specify their dominant moods. When inquired, they showed uncertainty as to confirm one or more as dominant mood(s) of their respective essays. When they were shown the results they confirmed the results being similar to what they intended to convey.

The results of seven essays which demonstrated entirely inaccurate psyches than the ones claimed by the writers called for the in depth analysis. All seven essays with the inaccurate results were analyzed individually to identify the reasons for inaccuracy in the reported results.

TABLE II. SUMMARY OF PSYCHEMAP RESULTS ACCURACY OF 50 ESSAYS

Accurate Results	**40%**
Results Close to Accuracy	**44%**
Inaccurate Results	**15.5%**

A total of 634 moods-related words were interpreted based on their intensity. As the Table above suggests, the 40% accurate and 44% close-to-accurate results of sample texts (50 essays written by non-native undergraduate university students of English language) show that the overall psyche of a text can be accurately determined on the basis of moods interpretation based on their intensity. It has been proved that if a writer uses moods-related-words with the exact intended expressions shared by the same non-native English language users, the interpretation of those moods-related-words can be accurate.

The 44% close-to-accurate results indicate the right interpretation of the expression of moods but not the exact determination of the overall psyche of a text. These results show that the writers in this case, though chose appropriate moods-related-words but the interpretation in order to be accurate requires that they should have more knowledge of the intensity of intended moods-related-words. The key to accuracy is understanding of the intensity of a psyche word as to enable a writer to control its occurrences and should use its related words accordingly.

The 15.5% inaccurate results have opened many areas for further study. The writers of 7 out of 50 essays demonstrated lack of linguistic competence with regard to the appropriate use of moods-related-words but also cohesiveness in the flow of thoughts in their writing composition. One of the areas in which the study can be further expanded in case of inaccurate results is to investigate that if these writers use the text mining tool as a learning device, then, what level of improvement can be achieved in making their expression of thoughts more cohesive and intended psyche accurate.

The inaccurate results have also unfolded complexities which are involved in the interpretation of multiple emotions within the same text to determine the accuracy of the extracted psyche. It has been observed that the writers have been astray in the expression of multiple moods and could not themselves specify their intended dominant moods. Lack of command on English language is seen to be a strong factor in their performance.

7. CONCLUSION

The study has showed that when specific extracted moods information was transformed into numbers for mining process, the overall text psyche was accurately determined. The results show that the study has been rightly attuned to tag local psyche on account of its basis in determination of moods, their synonyms and their respective weights, which was undertaken locally from non-native English language teachers who teach at undergraduate level. It has been proved that if a writer uses moods-related-words with the exact intended expressions shared by the same non-native English language users, the interpretation of those moods-related-words can be accurate.

REFERENCES

Bartell, B.T., Cottrell, G.W., And Belew, R.K. 1992. Latent semantic indexing is an optimal special case of multidimensional scaling. Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 161-167.

Budanitsky, A. And Hirst, G. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA.

Church, K.W., And Hanks, P. 1989. Word association norms, mutual information and lexicography. Proceedings of the 27th Annual Conference of the Association of Computational Linguistics. Association for Computational Linguistics, New Brunswick, NJ, 76-83.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., And Harshman, R. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics,19, 61-74.

Firth, J.R. 1957. A Synopsis of Linguistic Theory 1930-1955. In Studies in Linguistic Analysis, Philological Society, Oxford, 1-32. Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, Longman, London, 1968.

Gilad Mishne, I. I. Experiments with mood classification in blog posts. (www.wikipedia.org) (http://www.dcc.uchile.cl/~rbaeza/handbook/text_a.html)(http://portal.acm.org/citation.cfm?id=588074&dl=GUIDE)

Hanks., K. C. 1989. Word association norms, mutual information and lexicography. In Proceedings of the 27th Annual Conference of the ACL, New Brunswick, NJ.

Hatzivassiloglou, V., and Mckeown, K.R. 1997. Predicting the semantic orientation of adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th

Conference of the European Chapter of the ACL. Association for Computational Linguistics, NewBrunswick, NJ, 174-181.

Hatzivassiloglou, V., and Wiebe, J.M. 2000. Effects of adjective orientation and gradability on sentence subjectivity. Proceedings of 18th International Conference on Computational Linguistics. Association for Computational Linguistics, New Brunswick, NJ.

Hearst, M.A. 1992. Direction-based text interpretation as an information access refinement. In P. Jacobs (Ed.), Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. Lawrence Erlbaum Associates, Mahwah, NJ.

Hovy, S.-M. K. Determining the Sentiment of Opinions. In Proceedings of the 20th International Conference on Computational Linguistics (COLING), 2004.

Kamps, J., And Marx, M. 2002. Words with attitude. Proceedings of the First International Conference on Global WordNet, CIIL, Mysore, India, 332-341.

Landauer, T.K., And Dumais, S.T. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T.K. 2002. On the computational basis of learning and cognition: Arguments from LSA. To appear in B.H. Ross (Ed.), The Psychology of Learning and Motivation.

Littman, P. T. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4), 2003.

McKeown, V. H. Predicting the semantic orientation of adjectives. In Proceedings of ACL- 97, 35th Annual Meeting of the Association for Computational Linguistics, pages 174–181, Madrid, ES, 1997.

Osgood, C.E., Suci, G.J., And Tannenbaum, P.H. 1957. The Measurement of Meaning. University of Illinois Press, Chicago.

Pang, B., Lee, L., And Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 79-86.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. roceedings of the 14th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, 448-453.

S. D. Durbin, J. N. A system for affective rating of texts. In Proceedings of OTC-03, 3rd Workshop on Operational Text Classification, Washington, US, 2003.

Sarfraz, S. 2010. Accuracy of Text Psyche Based on Moods Interpretation, VDM Verlag Dr. Müller. ISBN 978-3-639-23547-0

Sholom M.Weiss, N. I. (2005). Text Mining. New York: Springer.

Spertus, E. 1997. Smokey: Automatic recognition of hostile messages. Proceedings of the Conference on Innovative Applications of Artificial Intelligence. AAAI Press, Menlo Park, CA, 1058-1065.

Turney, P.D. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the

Twelfth European Conference on Machine Learning. Springer Verlag, Berlin, 491-502.

Turney, P.D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the Association for Computational Linguistics 40th Anniversary Meeting. Association for Computational Linguistics, New Brunswick, NJ.

Wiebe, J.M., Bruce, R., Bell, M., Martin, M., & Wilson, T. 2001. A corpus study of evaluative and speculative language. Proceedings of the Second ACL SIG on Dialogue Workshop on Discourse and Dialogue.