Using Assorted Dataset For Emotion Analysis In Text

K. Jayanthi[1],Dr.D.Kavitha[2]

[1] Full Time Research Scholar , Department Of Computer Science And Applications, St. Peter's Institute Of Higher Education And Research, Chennai, India.

[2] Associate Professor, Department Of Computer Science And Applications, St. Peter's Institute Of Higher Education And Research, Chennai, India.

Jayanthi.Sharathkumar@Gmail.Com

Abstract

Based On The Machine Learning Approach This Paper Acknowledges The Emotions In Six (Anger, Disgust, Fear, Happiness, Sadness And Surprise) Employing A Contrasting Emotion-Annotated Dataset Which Mixes Fairy Tales, Blogsand Headlines Of News. Various Components Set Such As,Bags Of Words, And N-Grams, Were Used For This Purpose. The Technique Supportsin The Paper Is The Process Of Vector Machines Classifier (Vmc) Executed Considerably Better Than Other Classifiers, And It Generalized Well On Unobserved Examples.

1 Introduction

Nowadays In Many Research Areas The Emotional Aspects Attract The Eye Of The Many Researchers, Not Only In Engineering, But Also In Communication, Healthcare, Psychology, Etc. For Example, In Healthcare How The Diseases In The Brain Acquiredis Obtained As The Interest Of Some Researchers (E.G., Parkinson) That Affect The Power To Speak Emotionally. Otherwise, With The Affective Computing Emergency Level Within The Late Nineties, Numerous Technology Areas Of Several Researchers, E.G., Human Computer Interaction (Hci), Natural Language Processing (Nlp), Etc. Are Widely Interested In Emotions. Their Target Is To Advance The Machines That May Determineend Users' Emotions In Text And Direct Different Forms Of Emotion. The Highest Essential Way For A Computer To Recognize The User Emotions Automatically Is To Detect His Emotion From The Text That He Involves In The Entered Text Which Can Be A Blog, In A Web Chat Side Or In Any Other Type Of Text.

Generally, There Are Two Approaches The First Approach Is Based On The Knowledge-Based And The Second Approach Is The Machine Learning Were Adopted To Analyse The Emotions In The Text Automatically, About To Detect The Writer's Emotional State In The Text. The Primary Approach Consists Of Using Linguistic Models Or Knowledge In Prior To Classify Emotional Text. The Second Approach Is Used On The Supervised Learning Algorithms Which Makes The Models From Corpus Annotation. In The Sentiment Analysis, The Technique Which Is Based On Machine Learning Tends To Get Improved Results Than The Technique Of Lexical-Based, Because They Will Adapt Well To Distinct Domains. In This Research Paper, We Approvethe Approaches Of Machine Learning To Recognize The Emotions From Text. For This Objective, The Heterogeneous Dataset Used Here Is Collected From Headlines Of News,Blogs And Fairy Tales.

The Rest Of The Paper Is Organized As Follows: Section 2 Identifies The Several Datasets That We Used For Our Emotion Detection In Text. In Section 3, We Describe The Using A Heterogeneous Dataset For Emotion Analysis In Text 63 Methodology That We Adopted For This Purpose. Section 4 Presents And Discusses The Results By Comparing Different Machine Learning Techniques For Detecting Emotion In Texts. Finally, Section 5 Concludes The Paper And Outlines The Future Direction Of Our Research.

2 Datasets

In This Paper, We Reported The Experiments By Using Five Datasets. We Describe Each One In Details Below.

2.1 Text Affect

This Information Is Derived From The Headlines Of The Newspapers And Also Pulled It From The Google Search Engine And It Consist Of Two Sections. First Part Is Established With 250 Annotated Sentences Which For The Training Purpose And Another Part Is Established With 1000 Annotated Sentences Which Is For The Testing Purpose. Emotions Are Classified Into Six Categories Which Are Disgust, Fear, Sadness, Anger, Surprise As Well As Joy. Annotate Sentences Are Developed By The Degree Of These Emotional Load. Each Emotion Is Represented By Vector Of Scores, Instead We Can Label The Dominant Emotion As Sentence Label For Our Experiments.

2.2 Neviarouskaya Et Al.'S Dataset

There Are Two Datasets Found By The Authors In These Experiments. Three Annotators And Ten Labels Are Expected To Annotate Sentences In These Datasets. Labels Needs To Deal With Nine Emotional Types Defined By Izardand A Neural Type, These Emotions Are Disgust, Shame, Sadness, Joy, Interest, Guilt, Fear, Anger And Surprise. Sentences With Is Agreed More On Emotional Types Or Two Annotators Are Only Considered In These Experiments.

Following Are The Brief Description Of The Two Datasets:

- Dataset 1

This Dataset Has 1000 Sentences Derived In 13 Diverse Types(Wellness, Education And Health Etc,) From Different Stories.

- Dataset 2

This Dataset Derived From Diary- Like Blog Post Collection Which Has 700 Sentences

2.3 Alm's Dataset

These Datasets Are From Fairy Tales, The Data Is Included For The Sentence Annotations. The Sentences We Used For Our Experiments Is With High Annotation Agreements, In Other Words The Emotion Labels Identically Over The Sentences. There Are Emotions Which Classified Into Five From The Ekman's List Such As Surprised, Sad, Happy, Fearful, Angry-Disgusted And Angry-Disgusted Were Used For Sentences Annotations. Considering The Related Semantics And The Sparsity Of Data Between Disgust And The Anger, Both Of These Emotions Were Combined The Datasets Simultaneously By The Writer, To Stand As One Class.

2.4 Aman's Dataset

These Datasetsare Collected From Blogs Consists Of Emotion-Rich Terms [3]. These Sentences Are From Four Annotators Labelled With Emotions. The Terms Which The Annotators Agreed Are Only Considered On This Emotional Category. According To Ekman's Basic Emotions Are Categorized Such As Happiness, Sadness, Anger, Disgust, Surprise, And Fear And Also A Neutral Categorywas Used For Sentences Annotation.

3 Emotion Detection In Text

Weka Software [14] With Bow Representation Are Used For Comparison Of Three Classification Algorithms And It Gives Effective Result For Emotion Analysis In Text. Three Classification Algorithms Are: J48 For Decision Trees, Naïve Bayes For The Bayesian Classifier And The Smo Implementation Of Vmc. Quality Emotional Classification Of Text Will Be Feasible Only By Choosing The Required Effective Features Sets.

Below Provided Experiments Are Implemented Real Time. It Follows:

Bag-Of-Words (Bow) – Here Each And Every Sentence In The Dataset Comprised By A Feature Vector Composed Of Boolean Attributes For Each Word That Takes Place In The Sentence.For Each And Every Word Occurred In The Sentence, It Maps Actual Attribute Will Set To 1 Or Else It Will Be Set To 0. In Bow Each Word Is Consider As Independent Entities And It Won't Take Into Account Any Semantic Error Information Out From Text. Hence, Itprovides Us Usually Perfect One In Text Classification.

N-Grams – Here Sequentially Words Are Represented As Per Length N. N-Grams Helps To Find Out Syntactic Patterns In Text And Additionally It Highlights The Text Features Like Negations, E.G., "Unhappy Day". Negation Is The Method Of Analysing And Representing The Sentence In An Expressed Emotion And It Differentiate The Text Totally And Adds More Flavour To The Sentence. For Instance, If Sentence Has "It's Unhappy Day" It Should Be Analysed And Differentiated Into The Sadness Category Not Into A Happiness Category. For These Extensive Usages, Some Research Studies In Sentiment Analysis Claimed That N-Grams Performs High Level Of Sentence Comparatively Better Than The Bow Approach [4].

Lexical Emotion Features – Here This Method Features Represents The Set Of Emotional Words Extracted From Affective Lexical Repositories Such As, Wordnet Affect [13]. We Used In Our Experiments All The Emotional Words, From The Wordnet Affect (Wna), Associated With The Six Basic Emotions.

4 Results And Discussion

For An Exploratory Purpose, We Conducted Several Experiments Using The Labelled Datasets For Classifying Emotional Sentences.

4.1 Cross-Validation

For Proper Emotional Sentence Classification, It Is Important To Prepare The Data First. For Categorizing Text Into Emotional Classification, Individual Words Like "I" And "The" Are Visibly Not Useful And It Should Be Removed As Well. Furthermore, In Order To Decrease The Words Which Are Large In Numbers In The Representation Of Bow Used The Technique Of Stemming Lovinsstemmer From The Weka Tool, Which Displace A Word By Its Stem.

The Other Extensive Method To Reduce The Words Numbers In The Representation Of Bow Is To Increase The Negative Long Forms By Decrease The Negative Long Forms And Replace By Long Forms, For Example., Theword "Don't" Is Restored By "Do Not", Also The Word "Should Not" Gets Replaced Instead Of "Shouldn't" And So On. On Trying This Process Of Normalizing Negative Forms Gave Us Exceptional Results For The Representation Of Bow And In N-Grams The Effective Negative Expressions Are Considered. Later On, The Features Include Trigrams, Words And Bigrams.

In These Principles Of Exploration, To Train The Supervised Machine Learning Algorithms We Used Five Types Of Datasets Such As, The Global Dataset, Aman's Dataset And The Text Affect, And Last The Alm's Dataset. From Weka Software As A Baseline, We Also Used The Zeror Classifier; In The Training Set It Ranged The Data Into The Most Frequent Class.

Table 1. Outcomes Of Using Accuracy Rate (%) For The Training Datasets

	Naive Bayes	J48	Baseline	Smo

Global Dataset	39.6	32.8	31.6	39.6
Aman’s Dataset	54.92	47.47	36.86	61.88
Text Affect	73.02	71.43	68.47	81.16
Alm’s Dataset	59.72	64.70	50.47	71.69

The Results In The Table Shows The Highest Accuracy Rate For Each Dataset By Using The Smo Algorithm. In The Training Global Dataset Used Is Much Better Because On One Side It Contains Contrasting Data Which Are Gathered From The Headlines Of The News, Fairy Tales And From Blogs And On The Other Side The Variance Between Accuracy Rates Between Baseline And For The Smo Algorithm Is Highly Compared To Datasets Of Aman’s. When Compared To The Next Best Classifier The Smo Algorithm With The Global Dataset Is Statistically Better With A Best Level Of Confidence Of 95% Based On The Accuracy Rate.

Especially, We Have Reached An Exact Rate Of 81.16% For Aman’s Dataset Is Better Than The Accuracy Rate Which Is The Highest Of 73.89% Reported In [2]. Approximately Compared To The Work, We Not Only Used The Emotional Terms Of Words But Also The Emotion Less I.E., Non-Emotional Ones, As We Believe That We Can Express The Emotions Through The Underlying Meaning Of Some Sentences And The Contexts That Depends, I.E., “Thank You So Much For Everyone Who Came”. From The Sentence We Can Understand That It Does Not Indicates Any Word Emotionally But It Expresses The Happiness.

4.2 Supplied Test Set

On Presenting The Execution On The Training Datasets, The Extensive Issue That We Need To Observe In Analysing The Emotions In The Text Is To Deduce The Capacity On Unseen Examples, Since It Depends On The Vocabularies And The Context In Sentences Are Used. Thus, Using These Three Kinds Of Feature Sets (Bow, N-Grams, Emotion Words From Wordnet Affect) We Tested Our Model With Three Datasets(On The Trained Global Dataset). Table 2 Represents The Results.

Table 2. Smo Results Using Different Feature Sets

Test Sets	Feature Sets	Accuracy Rate (%)	
		Baseline	Smo
Text Affect	Wna	37.26	36.28
	Bow		39.38
	Bow +Wna		32.85
	N-Grams		**40.89**
Neviarouskaya Et Al.’S Dataset 1	Wna	25.72	45.66
	Bow		**58.84**
	Bow +Wna		55.25

	N-Grams		48.52
Neviarouskaya Et Al.’S Dataset 2	Wna	34.88	49.25
	Bow		**58.45**
	Bow +Wna		57.24
	N-Grams		50.85

As Represented In Table 2, Using The Representation Of N-Grams For Text Affect Lay Out The Comparatively Better Results Than The Representation Of Bow, But The Difference Is Not Statistically Significant.

Conclution:

This Paper Represents The Emotional Recognition Of The Text Using Supervised Algorithms On Datasets Automaticallyby Using Machine Learning Method Approach.To Achieve This, We Use Diverse Dataset Collection From Blogs, Fairy Tales And From Headlines Of News And Compared Each Contrasting Dataset Into A Separate Training Data. Furthermore, To Generalize The Unseen Examples The Smo Algorithm Approaches Well And Showed That The Algorithm Made The Improvement Significantly Better Than The Other Classification Algorithm.

References

1. Alm, C.O.: Affect In Text And Speech. Phd Dissertation. University Of Illinois At Urbanachampaign (2008)
2. Aman, S., Szpakowicz, S.: Identifying Expressions Of Emotion In Text. In: Matoušek, V., Mautner, P. (Eds.) Tsd 2007. Lncs (Lnai), Vol. 4629, Pp. 196–205. Springer, Heidelberg (2007) Using A Heterogeneous Dataset For Emotion Analysis In Text 67
3. Aman, S.: Identifying Expressions Of Emotion In Text, Master’s Thesis, University Of Ottawa, Ottawa, Canada (2007)
4. Arora, S., Mayfield, E., Penstein-Ros, C., Nyberg, E.: Sentiment Classification Using Automatically Extracted Subgraph Features. In: Proceedings Of The Naacl Hlt 2010 Workshop On Computational Approaches To Analysis And Generation Of Emotion In Text (2010)
5. Ekman, P., Friesen, W.V.: Facial Action Coding System: Investigator’s Guide. Consulting Psychologists Press, Palo Alto (1978)
6. Izard, C.E.: The Face Of Emotion. Appleton-Century-Crofts, New York (1971)
7. Melville, P., Gryc, W., Lawrence, R.: Sentiment Analysis Of Blogs By Combining Lexical Knowledge With Text Classification. In: Proc. Of Kdd, Pp. 1275–1284 (2009)
8. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: @Am: Textual Attitude Analysis Model. In: Proceedings Of The Naacl Hlt 2010 Workshop On Computational Approaches To Analysis And Generation Of Emotion In Text, Los Angeles, Usa (2010)
9. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Compositionality Principle In Recognition Of Fine-Grained Emotions From Text. In: Proceedings Of The International Conference On Weblogs And Social Media. Aaai, San Jose (2009)

10. Paulmann, S., Pell, M.D.: Dynamic Emotion Processing In Parkinson's Disease As A Function Of Channel Availability. Journal Of Clinical And Experimental Neuropsychology 32(8), 822–835 (2010)
11. Picard, R.W.: Affective Computing. Mit Press, Cambridge (1997)
12. Strapparava, C., Mihalcea, R.: Semeval- 2007 Task
13. 14: Affective Text. In: Proceedings Of The 4th International Workshop On The Semeval 2007, Prague (2007)
14. Strapparava, C., Valitutti, A., Stock, O.: The Affective Weight Of Lexicon. In: Proceedings Of The Fifth International Conference On Language Resources And Evaluation, Genoa, Italy (2006)
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools And Techniques, 2nd Edn. Morgan Kaufmann, San Francisco (2005)