Improved Sentiment Classification From Meeting Transcripts

J.I.Sheeba[1] and Dr.K.Vivekanandan[2]

[1]Assistant Professor,
Department of Computer Science & Engineering,
Pondicherry Engineering College, Puducherry, India.

[2]Professor,
Department of Computer Science & Engineering,
Pondicherry Engineering College, Puducherry, India.

Abstract

The web provides volumes of text-based data about customer preferences which are stored in online review websites, twitter, face book, blogs, etc. Sentiment classification has emerged as a method for mining opinions from such text archives and it uses machine learning methods combined with linguistic attributes or features in order to identify the sentiment polarity like positive, negative, and neutral for a particular document. Topic detection is also considered which helps in detecting the sentiment of each topic. In this proposed framework, it is able to identify sentiment and topic and also author classification from meeting transcripts. By improving its performance, it will incorporate the features of both JST and FRN method used here, it can identify sentiment and topic from the transcripts. In this Author classification, it is possible to identify both Author identification and Author characterization. Using SVM classifier, sentiments, topic and author classification are extracted .

Keywords: *Sentiment classification, Topic detection, Author characterization, Author classification, Author identification*.

1. Introduction

Now a days, the vast amount of text data is available online, text mining has been applied to discover hidden knowledge from text in many applications and domains. In the business field, it is mainly used to find out customers, sentiments and opinions about company's products and services. Manually discovering sentiments and opinions from large volume of data is very difficult .Also ,another issue of this problem is to extract the customer's opinions from large volume of unstructured text data (Ex :Meeting transcripts).

Sentiment analysis is also known as sentiment classification or opinion mining . It is a computational technique that seeks to understand and explain opinions and sentiments by analyzing large amounts of opinion data in such an efficient way as to assist in human decision making. It is essentially useful for all the fields in our day to day life. For example ,in the business field, sentiment classification can help companies to analyze customer's opinions for improving their products, providing better customer service and identifying new business opportunities. In politics, it can predict to shifts in public opinion regarding election candidates and also it will give the suggestion to select the candidate. In daily life, it suggests the people to select electronic products like mobile, camera etc, it will give better decision to see the movies or books to read, and to buy a product[1].

In this paper the domain meeting transcripts has been focused. Meeting speech is significantly different from written text and other speech data. For example, in meeting transcripts many people can participate, even the discussions are not organized well, and the speech is unplanned one and it contains disfluencies and also the sentences are not constructed well .The people who are involved in the meeting speak different pronunciations and they use different types of words. Each people can act as different roles and topics in the transcripts. So extracting keywords from meeting transcripts is difficult one compared to the documents[2]. Keyphrases as a brief summary of a document provide a solution to help organize, search and retrieve documents very fast. Keyphrases are also called as Noun phrases (NPs) that can reflect the main content of documents. Keyphrase extraction method aims to select a set of terms like bigram, trigram and N-gram words from a given document .It is used for various fields like Natural Language Processing (NLP) applications such as summarization and Question-Answering (QA) and search engines field , full-text indexing and assist users creating good queries[3].

In this framework ,N gram features are going to identify using FRN method. In this proposed framework it will classify the sentiments from meeting transcripts and also identify topic ,author classification in a single framework it self.

2. Related Works

In Sentiment Classification method (Chenghua Lin, Yulan He ,2011)proposed unsupervised framework called Joint Sentiment-Topic (JST) model based on Latent Dirichlet Allocation (LDA). JST detects sentiment and topic simultaneously from text compared to existing approaches and it performs well[4].In another method(Giacomo Inches,2011) proposed the ideas of author characterization, author identification and also additionally identify the Topic. Using these three concepts it is easy to detect the people who are involved in the conversation, find out author character and topic of particular transcripts[5]. Another approach (Ahmed abbasi,2011)proposed Feature Relation Network for Sentiment Classification ,it will identify only N-gram based sentiments. In this method they introduced subsumption, parallel relation concepts for reducing redundant words[6].In (Yulan He Chenghua Lin,2011)proposed a modified Jst model by incorporating word polarity priors through modifying the topic-word Dirichlet priors.The polarity bearing topics are extracted by Jst model[7].

In (Xuan-Hieu Phan,2011) proposed a Hidden Topic-Based Framework with Short Web Documents using LDA model. Here,they solved two problems like data sparseness ,synonyms problem using LDA method through MaxEnt classifier.LDA method is used to identify the low frequency keywords and topics[8]. In (Alexandra Balahur ,2012) proposed to detect implicit expression from text. In the Most existing approaches based on word-level analysis of texts and are mostly able to detect only explicit expressions of sentiment. The future aim of this is paper is to detect implicit expression from text using JST model[9]. In keyword Extraction method (Feifan , Fei liu 2011) proposed Single-loop feedback strategy for keyword extraction using some additional features like traditional frequency or position-based clues, term specificity features, decision-making sentence-related features, and group of features derived from summary sentence. In this paper data set meeting transcripts has been taken[10].
In this proposed model, it is going to incorporate the features of both JST&FRN model for sentiment classification and identify the topic and also it includes the author identification and characterization.

3. Overall Architecture

The investigation of sentiment classification on meeting domain is done. This process is done step by step for the best result in extraction of sentiment process. In this unsupervised approach, FRN method is used to classify the sentiment and JST method which is used for identifying both sentiment and topic and also it includes another feature for author classification which is nothing but author detection and author characterization. Author identification is used to detect the people who are involved in the conversation and characterization of author which is to group the author depending on their sentiment. Comparison of JST, FRN , Human Annotation methods is done. SVM classifier is used to extract the topic and sentiment from meeting transcripts.

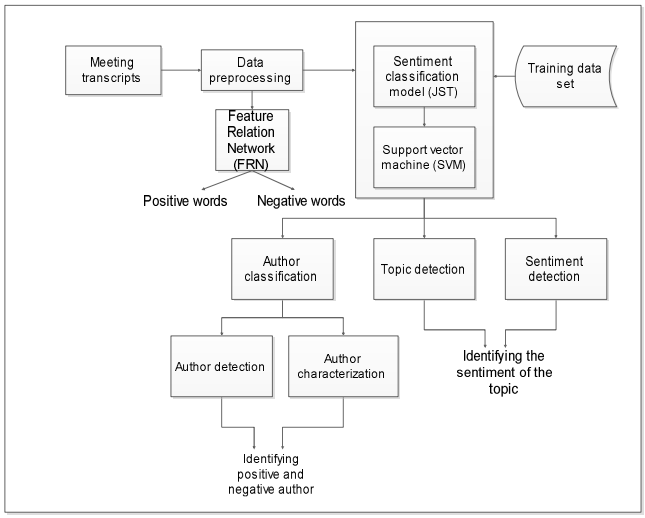


Figure 1 Proposed UnSupervised Framework for Sentiment Classification from Meeting Transcripts

The figure 1 shows a Unsupervised frame work for Sentiment Classification for given transcripts , with data flow along the arrows. In this task, 4 Steps have been included

1. Sentiment classification from meeting transcripts using FRN
2. Sentiment classification and Topic detection from meeting transcripts using JST
3. To Detect Author classification(Identification, Characterization) from the Transcripts
4. Sentiment classification and topic extraction through SVM classifier

3.1 Sentiment classification from meeting transcripts using FRN

3.1.1 Speech to text

Speech to text is translation of audio voice into text form automatically. Speech to text is otherwise known as automatic speech recognition or voice recognition or computer speech recognition. Here, the term "voice recognition" is sometimes used to refer to recognition of systems that must be trained to a particular speaker in the desktop recognition software. Here, Dragon naturally speaking software provides better accuracy compared to other recognition software and it has been used to provide the powerful software for converting voice to text which is used here.

3.1.2 Data Preprocessing

Data preprocessing is mainly used for removing irrelevant and redundant information present or noisy and unreliable data in the meeting transcripts. It includes cleaning, selection and feature extraction , normalization, transformation etc . This paper has undertaken stop word removal process for data pre-processing.

Stop words are common words that contain less important information than keywords. Stop words must be removed in order to save memory space or to speed up search results. As the list of stop words cannot be removed automatically and it must be trained by human input.
Examples of Stop words: A, About, Being, Can, Ever, of, the ,You etc.

3.1.3 Feature Relation Network (FRN)

A Rule-based multivariate text feature selection method is called Feature Relation Network[6] that considers semantic information and also leverages the syntactic relationships between N-gram features. Feature Relation Network is intended for the inclusion of extended sets of heterogeneous N-gram features for enhanced sentiment classification .It includes the following steps for classifying the sentiments like N-Gram feature generation, Feature selection, Feature extraction and Relation between features.

N-gram based Feature Extraction

Keyphrases are the combination of 2 or more words which describe a meaningful and important content in a document. Only minimal documents have the author assigned keyphrases. Extracting the keyphrases manually is a difficult task. So, it must be automated. Keyphrases give the high-level description of the documents content and it is mainly used for users to decide whether the particular document is relevant or not.

Keyphrases summarize documents very quickly and it can be used as a low-cost measure of similarity between documents, additionally, it is used to cluster documents into groups .If the keyphrase is entered into a search engine , all documents with this particular keyphrase attached will be returned to the user[11].

The keyphrases are Noun phrases which represent the main content of the documents. Noun phrases describe the meaningful phrases and it may be used in fields like intrusion detection, quality of service ,text summarization etc. Keyphrases can be simple words or combination of 2 or more words. It may also contain hyphens(e.g sensor grouping) and apostrophes (e.g Bayes' theorem).The part of speech (POS) is to be assigned for each word in the document and if the words in the document have 2 and 3 nouns consecutively and it is taken as bigram and trigram respectively. Here N-gram features are extracted by FRN method.

The most popular class of features used for opinion mining is N-grams . N-gram is a subsequence of n items from a given sequence and the items can be phonemes, words, syllables, letters, base pairs according to the application. N-gram features can be classified into two categories: fixed and variable. Fixed N-grams are exact sequence occurring at either the character or token level and Variable N-grams are extraction patters capable of representing more sophisticated linguistic phenomena. A fixed and variable N-grams have been used for opinion mining, including word, POS ,syntactic, character, legomena, and semantic N-grams. Word N-grams include bag-of-words and higher order word N-grams (e.g., bigrams, trigrams).

Feature Extraction

Dictionary is created and weights are assigned to each word. The weights for each word are calculated from sentiwordnet. It is a source available where each particular word has been assigned a positive and a negative score. The generated N-gram words are compared with the initial feature set database. The words which match with the database alone are retrieved.

Feature Selection

Feature selection method is mainly used to improve classification accuracy, and provide greater insight into important class attributes. Feature selection methods can be divided into two types univariate and multivariate.

Univariate methods consider attributes individually, computationally efficient, evaluating individual attributes .This method has one disadvantage that is important attribute interactions have not been considered. Multivariate methods consider attribute groups or subsets and also it considers attribute interactions. FRN uses two relatios like subsumption and parallel relations .It is mainly used to remove redundant and irrelevant N-grams in the meeting transcripts.

Subsumption Relations

A subsumption relation occurs between two N-gram feature categories such as one word subsumes with other word. For example, word unigrams subsume word bigrams and trigrams, word bigrams subsume word trigrams. For example, in the sentence "I LOVE INDIA" there are six word N grams: I, LOVE, INDIA, I LOVE, LOVE INDIA, and I LOVE INDIA. Here the unigram word LOVE represents positive sentiment.

Parallel Relations

A parallel relation occurs where two heterogeneous same order N-gram feature groups may have some features with similar occurrences. Here unigrams word (1-Word) can be associated with many POS tags (1-POS) and vice versa. For example, the POS tag ADMIRE_VP and the semantic class SYN-Affection both represent words such as "joy" and "happy." If A and B are considered two words , A is considered to be parallel to B (A—B) .if two words are giving same meaning any one of the words will be considered and extracted from the transcripts and other word will be removed ,by this way it is able to reduce the redundancy keywords from the transcripts[6].

3.2 Sentiment classification and Topic detection from meeting transcripts using JST

3.2.1 Joint Sentiment-Topic (JST) Model

Joint Sentiment-Topic (JST) model is mainly used to detect sentiment and topic simultaneously from text. It is mainly based on Latent Dirichlet allocation (LDA) method. Some sentiment classifications methods are often fail to produce satisfactory performance when shifting to other domains, the nature of JST makes it highly portable to other domains. The LDA model has three hierarchical layers, where topics are associated with documents, and words are associated with topics. But in the Joint Sentiment-Topic (JST) model, an additional sentiment layer has been added between the document and the topic layer. JST is basically a four-layer model, here sentiment labels are associated with documents and topics are associated with sentiment labels and words are associated with both sentiment labels and topics. In this proposed un supervised framework JST model is used to detect topic and sentiment simultaneously from meeting transcripts. Results are compared to FRN method and select more efficient one[4] .

3.2.2 Topic Extraction

JST extracts the topics and evaluates the effectiveness of topic sentiment captured by the model. Topics extracted from only one dataset under positive and negative sentiment labels. The classifier classifies the dataset and the data set compared with training data set for identifying the topic. Topic identification using SVM classifier for the best accuracy from the conversation [4].

3.3To Detect Author Classification(Identification, Characterization) from the Transcripts

3.3.1 Author Classification

Author identification and author characterization are called author classification. To identify the author, POS tagger is used here.

POS tagger(Qtag tool)

The tool that reads text from meeting transcripts and for each token in the text returns the part-of-speech which is used for both stop word removal and author identification .

Author identification is used to identify the people who are involved in the conversation. It helps to identify the authors and their participants in the meeting transcripts. There are many authors involved in the meeting transcripts, here, the author has been identified with the help of POS(part of speech) concept.POS tag has been done for the meeting transcript using QTAG tool. It tags the each word's POS. There are some conditions to identify the authors

- The POS of the word must be NP (proper Noun)
- It should be the first word of a sentence.
- Next of the word must be a semicolon (:)

[Generally meeting transcripts author names are available in that format only]

If a particular word satisfies the above mentioned rules then the word will be selected as the name of the author of the transcripts .

Author characterization is used to identify the characteristics of each author and categorize the people involved in conversation. The sentiment of each author has been identified for detecting the attitude of the author.

Depending on the author identification, the meeting transcript spitted as each author's participation. Then the sentiment will be detected for every author using SVM classifier. Classification concepts are used to group the authors depending on their sentiment [5].

3.4 Sentiment classification &topic extraction through SVM classifier

A Support Vector Machine (SVM) is used for classification and Regression analysis. SVM takes a set of input data and predicts for each given input which of two possible classes comprises the input making the SVM a non probabilistic binary linear classifier. Here a set of training examples is given and marked belonging to one of two categories , SVM training algorithm builds a model that assigns new examples into one category or the other .

To train the SVM classifier and prepare the prediction model, some sample documents per each category are collected and each sample document is prepared by browsing the Web pages which are related with each category ,because each Web page is entitled by its topic and the category. Based on this predefined category newly inputted documents can be classified. The result from the SVM classifier is extracted as the number that stands for each category[12].

TFIDF kernel has been used to classify the text here .TFIDF Kernel is nothing but to run the bag of words approach and the string contents of each page has been parsed and assigned the weights according to the standard BOW (Bag of Words) formula

$$\text{weight(word)} = \text{TF}(w_i, document)\, x \log\left(\frac{n}{DF(w_i)}\right)$$

Here *TF(wi, document)* is the number of times the word *wi* occurs in the document, *n* is the total number of documents and *DF(wi)* is the number of documents in which the word *wi* occurs.
Corpus contains the predefined positive and negative category list which is used to train the SVM classifier and prepare prediction model. The test data is compared with trained data using SVM classifier and it classifies the positive and negative words. The positive words are labeled as +1 and negative words are labeled as -1 by SVM classifier[13].

4. Experimental Results

4.1 Data Set

The dataset used for this experiment is a meeting transcripts like T1, T2, T3, T4 and T5 and they are the inputs. The N-grams feature generation is used here for feature selection. The same meeting transcripts input is fed into JST and FRN and the number of features generated is listed in the table. The below table shows the comparison among number of features of human annotated, FRN and JST.

Table 1. Comparison of Human annotation, FRN and JST

INPUT	JST	Human Annotation	FRN
T1	15	12	8
T2	13	12	11
T3	22	22	15
T4	17	14	11
T5	6	6	5

From the results, it is clear that when comparing with three methods JST is the best for selecting feature selection.
The below figure shows the comparison among unigram and N-grams features. It is clear that from the results unigram features are more efficient than bigram. But when comparing with number of features, N-gram features are more in number.

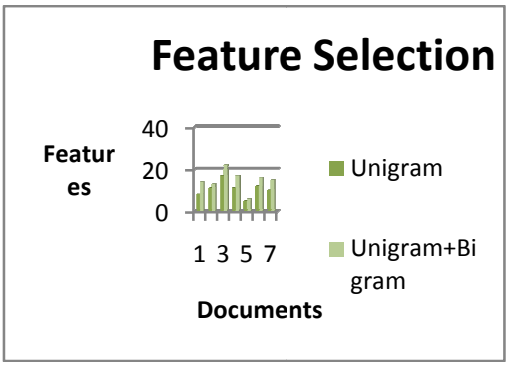


Figure 2. Feature Selection comparison: Unigram Vs Unigram+ Bigram

The below figure shows the comparison among Human Annotated (HA), JST and FRN. It is evident from the results that JST model gives more features.

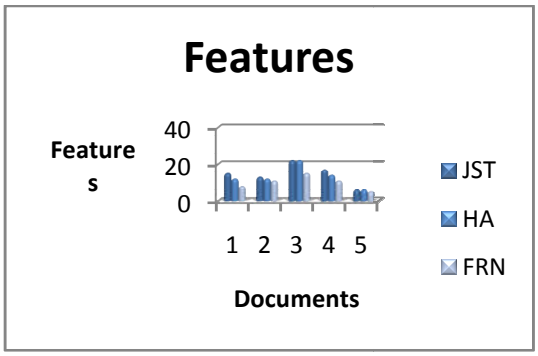


Figure 3. Feature comparison: JST Vs FRN Vs HA

Word weighting

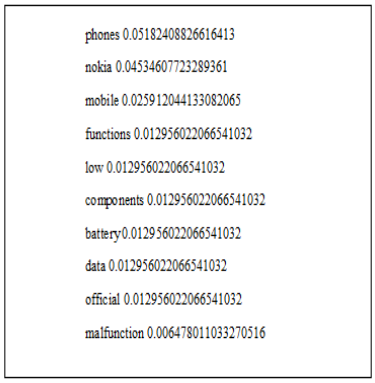
SVM classifier is used to detect the sentiment of the meeting transcripts.SVM deals with vector value. Vector represents weight of the words. The weight of the word will be calculated by using **tf-idf** concept. The **tf*idf** weight (term frequency–inverse document frequency) is a numerical statistic which reflects the importance of a word to a document in the collection.

Tf(t ,D) = Number of times t appears in that meeting transcripts D.

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tf} * \mathrm{idf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Sample results,



Topic sentiment and detection

Topic and sentiment have been detected in this model, given meeting transcripts words are compared with training set and then number of topics will be detected and sentiment detected for each transcript, finally topic's sentiment is assigned to each topic.

Table 2. Topic and sentiment detection

Meeting transcripts	Positive	Negative	No.of topics	Topic sentiment
T11	17	7	1 (Mobile)	Positive topic
T12	3	8	2(Book, Environment	Negative topic
T13	15	4	1 Research)	Positive topic
T14	12	7	1(Music)	Positive topic
T15	6	19	1(Politics)	Negative topic

Here input files are denoted as T11,T12,T13,T14,T15. SVM classifier is used to extract positive and negative word from the transcripts.

Author classification

Authors are the people those who are all involved in the conversation. POS concept is used to identify authors. And each author's sentiment was detected by using SVM classifier. Author's are classified depending on the author's sentiment. The result of the author classification is given below

Table 3. Author classification

Meeting transcripts	No.of. authors	Positive authors	Negative authors
T11	2	2(John,Peter)	NULL
T12	3	2(Arun,Charles)	1(Babu)
T13	2	2(Betty,Jenny)	NULL
T14	2	2(Ramu,Prabhu)	NULL
T15	3	1(Rahul)	2(Rajesh, Siva)

Here ,the topic, sentiment and author classification are combined together and done in a single framework. The result of the framework is represented in the below table and figures.

Table 4. Topic, sentiment and author classification

Meeting transcripts	Positive	Negative	No. Of topics	No. Of authors	No.of positive authors	No.of negative authors
T11	17	7	1	2	2	0
T12	3	8	2	3	2	1
T13	15	4	1	2	2	0
T14	12	7	1	2	2	0
T15	6	19	1	3	1	2

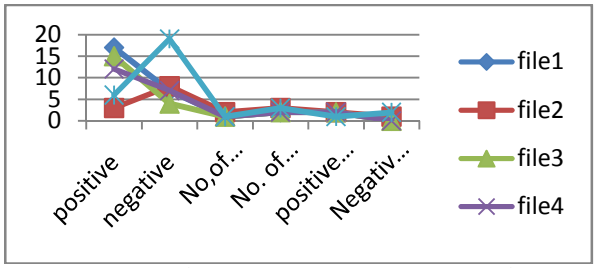


Fig 4. Graph for Topic, sentiment and author classification

5. Conclusion

This paper proposes a method for improved selection of text attributes for enhanced sentiment classification. It shows whether the given transcript is a positive or a negative oriented topic. Word sense disambiguation is also done which governs the process of identifying sensible words used in a sentence, when the word has multiple meanings. FRN's use of syntactic relation and semantic information regarding N-grams has enabled it to achieve improved results over various multivariate, univariate and hybrid feature selection methods. The results are proved in JST, most of the features match with human annotated than FRN. When comparison is done between JST outperforms and FRN. Author identification and characterization are also performed. Hence, it classifies the author and the topic sentiment and then it groups the author. It is performed in a single framework, so it reduces time and space complexity. It benefits to identify the members of the meeting transcripts and their opinion

In future, different methods can be applied and without using the training set, the system must be able to detect the topic and sentiments. The sentiment and topic identification approach can be extended to other languages. By adding some more additional features, this system can also perform automatic rating for online products.

References

[1] Zied Kechaou , Mohamed Ben Ammar "Improving e-learning with sentiment analysis of users opinions", 2011 IEEE Global Engineering Education Conference (EDUCON) "Learning Environments and Ecosystems in Engineering Education", Amman, Jordan,April 4 - 6, 2010.

[2] Feifan Liu, Deana Pennell, Fei Liu, "Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts", ACM ,2009.

[3] Su Nam Kim, Timothy Baldwin and Min-Yen Kan "Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction", Proceedings of the 23rd International Conference on Computational Linguistics , Beijing, August 2010, pages 572–580.

[4] Chenghua Lin, Yulan He, Richard Everson, and Stefan Rüger,"Weakly-SupervisedJointSentiment-Topic Detection from Text," IEEE Transactions on Knowledge and Data Engineering 07, Feb, 2011.

[5]Giacomo Inches, Fabio Crestani, "Online Conversation Mining for Author Characterization and Topic Identification" ,PIKM'11, October 28,ACM, 2011.

[6] Ahmed Abbasi, Stephen France, Zhu Zhang, Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," IEEE Transactions on Knowledge and Data Engineering, Mar. 2011,vol. 23, no. 3, pp. 44-42.

[7] Yulan He Chenghua Lin, Harith Alani "Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification". Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, June 19-24, 2011,pages 123–131.

[8] Xuan-Hieu Phan, Cam-Tu Nguyen, " A Hidden Topic-Based Framework toward Building Applications with Short Web Documents", IEEE Transactions On Knowledge And Data Engineering, VOL. 23, 2011.

[9] Alexandra Balahur , Jesús M. Hermida, "Detecting implicit expressions of emotion in text: A comparative analysis", Decision support system ,Elsevier, 2012.

[10] Fei Liu, Feifan Liu, and Yang Liu, "A Supervised Framework for Keyword Extraction From Meeting Transcripts",IEEE Transactions On Audio, Speech, And Language Processing, VOL. 19, NO. 3, March 2011,pp.538-548.

[11] Eibe Frank and Gordon W.Paynter, "Domain specific Keyphrase extraction".

[12]http://www.support-vector.net

[13] Veselin Stoyanov and Claire Cardie, "Topic Identification for Fine-Grained Opinion Analysis", International Conference on Computational Linguistics (Coling 2008), Manchester, August 2008, pages 817–824.