Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

# Facts or Feelings – A Novel Framework For Classifying Survey Text Data

**Debanjan Banerjee**[1]

[1]Dept. of MIS, SARVA SHIKSHA MISSION KOLKATA, Kolkata - 700042, India,

**Abstract -** The rapid growth of online survey responses on important social issues gives birth to the need for automated tools for analysing potentially beneficial patterns and trends for the society at large. In this paper we propose a framework to classify online survey responses as either intelligent or egoistic. We classify those survey responses as intelligent if they are based upon objective facts. Survey responses are classified as egoistic if these are based upon subjective considerations. We believe this approach will pave the way for future works to create a substantial approach for extracting knowledge out of textual data.

**Keywords** - Text classification, Survey text classification, Intelligent and egoistic classifications, Machine learning

## I. INTRODUCTION

The humongous growth in Internet communications has given rise to the phenomenon of online surveys. Increasingly people opine about their preferences and attitudes on various important social events such as sporting occasions and elections through online surveys. The sporting events such as the soccer world cup generate a lot of interesting views amongst these survey responders. Many marketing agencies can boost their marketing policy making once they are able to tap into these comments as this would be helpful to understand about popular mood over a well-known event. The primary challenge for entrepreneurs and researchers alike while dealing with survey reviews is to bring about organized frameworks that turn untidy and unstructured survey responses into meaningful information. Once the researcher community is able to come up with these frameworks it will enrich the efforts to understand how humans influence each other during popular social events through survey responses. The present work intends to aid this endeavour by classifying these survey responses into two distinct classes which are intelligent and egoistic. The survey questions have been designed with the objective of gauging popular opinion on the recently concluded World cup soccer in Brazil on June-July, 2014. This work will present a framework which will look into classifying the survey responses into two distinguishable classes i.e. intelligent and egoistic. This paper will first document in detail the past work which has been performed in the same area and will provide arguments to distinguish it from these past works. Then the work will detail the working process involved alongside the mathematical formulations which have been considered on the work. The results and the conclusions which follow from the results will also be documented later on.

## II. RELATED WORK

The presence of the sizable amount of online textual data has been an intriguing challenge for the researcher community for analysis and extracting useful knowledge out of the amorphous data. Many prominent and notable works have been pursued in this regard. Wiebe and Riloff et all have made significant contributions in terms of extracting subjective as well as objective sentences from unlabelled texts[1] , creating a bootstrapping method for extracting subjective nouns out of opinionated expressions [2] , creating high-precision classifiers for extracting subjective patterns out of opinionated expressions [3] , as well as extracting subjective nouns using bootstrapping methods [6]. Similar important contributions towards sentiment and polarity classification of online reviews.Pang and Lee described in detail about sentiment analysis [5], worked for identifying collocations for recognizing opinions [8]. Pang, Lee et al came up with machine learning methods for classifying sentiment in movie reviews [9], Yu et al worked with separating Facts from Opinions and Identifying the Polarity of Opinions [10] , Turney came up with a framework for semantic orientation applied to unsupervised classification of reviews [11] , Spertus et al worked for creating an automatic recognition of the hostile messages [12] , Das and Chen worked with opinion extraction for small talk on the web [13] Karlgren and Cutting [14] and Kessler et al [15] has worked upon text genre classification. The observable commonality in these pioneering works is that these efforts have been concentrated mostly upon the areas of subjectivity and objectivity classification, sentiment analysis, polarity classification, pattern extraction and genre classification.

Our work is fundamentally unique than these works in primarily two respects. This work concentrates upon classifying a survey response as either intelligent or egoistic. We are not aware of any other work in this regard before. Another unique aspect of our work is that we focus on the domain of online survey of soccer World cup 2014; whereas most other works deal with online reviews, newspaper articles and other versatile unlabelled textual sources. According to our research until now there has been no mentionable work in the past in either of the two domains.

## III. PROPOSED WORK

The work defines an intelligent survey response as the particular response which is expressed based upon objective facts and an egoistic response is the response which has been expressed by the responder with a subjective viewpoint.

The Cambridge dictionary of Psychology defines intelligence as “A set of abilities to adapt better to the environment through experience.”[4] This work considers this definition of intelligence as the basis of consideration and thus defines any survey response as intelligent which is expressed purely on factual or objective considerations.

By contrast we define any survey response as egoistic whereby the responder infers the survey question from an individual viewpoint and the response involves explicit subjective clues. To illustrate the classification let us consider two responses to the survey question of “Who will win the world cup soccer and why?” The first response is “Brazil will win as they are playing at home and they have won most number of world cups so far.”

Here the response can be considered as intelligent given our definition of the term since the response includes two observations “they are playing at home” and “they have won most number of world cups so far.” both of which happens to be factually correct in the context of the soccer world cup 2014. The response “Brazil will win as they are simply the best team in the world” is considered as egoistic in our work since this work contains explicit subjective terms “simply the best team in the world”.

The complete system has been built upon four basic unitary processes which work in congruence with each other. These are collection of survey responses, pre-processing of the survey data, manual annotation and finally application of the classifiers. The attribute selection procedure for classification works with multiple techniques which are attribute gain , gain ratio and symmetric uncertainty in order to determine the optimum set of attributes which give maximum accuracy.

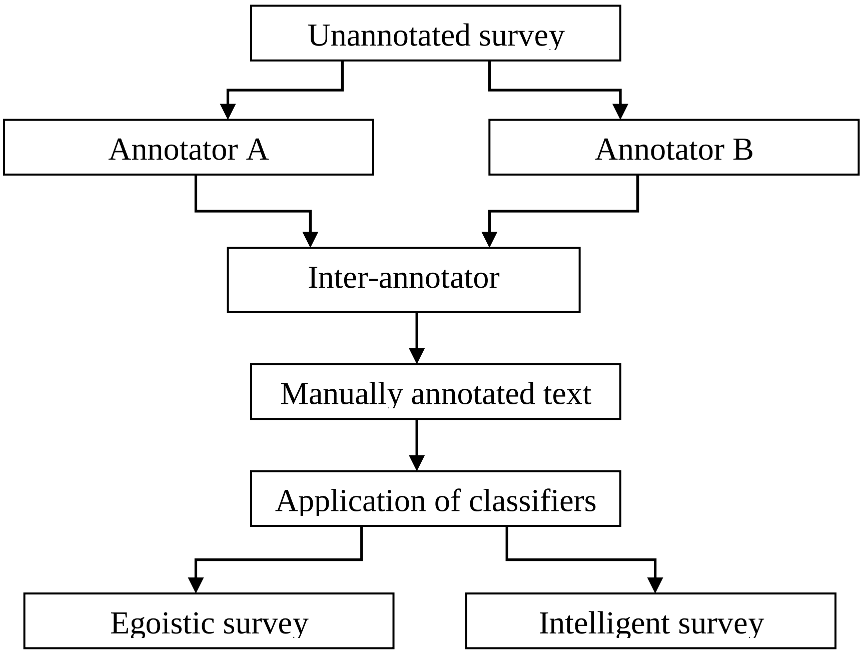

***Figure 1: The complete classification process of survey responses***

The overall process can be broadly categorized into four broad categories collection of survey responses, pre-processing of survey data, manual annotation and classification.

### 3.1 Collection of survey responses

The survey was set up based upon two specific questions. These questions were “Who do you think will win the World cup?” and “Why do you think so?” The survey was communicated to people chiefly through social media in general and the survey website surveymonkey.com in particular. The questions were taken throughout the duration of the world cup. The survey was concluded the very same day when the competition had ended.

The survey on the world cup was considered based upon certain assumptions. These assumptions determine our definition of what constitutes as intelligent or egoistic. The domain of the work consists on survey responses. The survey question was created based upon the assumptions that 1. Survey questions provide the responders the opportunity to express their views either from their personal viewpoint or from a factual and objective perspective. 2. The survey topic is on a subject which is a real world social event such as the world cup whose final outcome is yet not known to the user however once the outcome comes to reality it becomes an undisputable universal fact

### 3.2 Pre-processing of survey data

This process involves the purification of working data by removing survey responses which only provides responses to the question “Who do you think will win the World cup?” without giving any explicit answer to the survey question of “Why do you think so?” We consider such responses such as “England”, “brazil” or “Germany” as outliers to our work since these responses do mention the reason (whether from personal or subjective viewpoint) and henceforth these types of responses are not included in our classification process. The filtering out of these types of comments involves manual verification. Since world cup is an event which involves responses from people whose mother tongue is not English, we decided to consider only those responses which were presented in English.

### 3.3 Manual annotation

In this process two human annotators get involve with the process and annotate the responses as either egoistic or intelligent. The responses where there is an inter-annotator agreement are selected for applying classification techniques. The following confusion matrix presents the picture with regards to the inter-annotation agreement.

| | c | d |
|---|---|---|
| a | Total number of egoistic responses where there were agreement (x1) | Total number of egoistic responses where there were no agreement (x11) |
| b | Total number of intelligent responses where there were agreement (x2) | Total number of intelligent responses where there were no agreement (x22) |

So our purified data set consists of the combined total number of egoistic and intelligent responses where there was agreement between both the annotators. This value can be expressed as mutually agreed responses ~ x1 + x2

### 3.4 Classification

For our work each individual word belonging to the corpus is considered as an attribute and we have selected three types of attribute selection techniques which are Information gain, Gain ratio attribute and symmetrical uncertainty evaluation. These attribute selection techniques help the classification algorithms to consider the most relevant and promising attributes while filtering out the irrelevant and redundant attributes.

### 3.5 Information gain

The information gain attribute can be expressed as

**IG <- 1 – Entr**.

Whereas **Entr**. Represents entropy for each individual word .On the other hand Entropy can be represented as Entr. <- Probability of the individual word * Log (Probability of the individual word)

Probability of the individual word can be derived by the following ratio

Total number of occurrences of the individual word in the corpus / Total number of occurrences of all the words in the corpus

The work proceeds to prepare rankings based upon the information gain for all the attributes (the value is usually less than 1) to use this measure in classification techniques. The classification techniques which use this attribute are J48, random forest, K-nearest neighbour, naive bayes and SVM, respectively.

Let $Attr$ be the set of all attributes and $Ex$the set of all training examples, $value(x,a)$with $x \in Ex$ defines the value of a specific example $x$ for attribute$a \in Attr$, $H$ specifies the entropy. The information gain for an attribute $a \in Attr$is defined as follows:

$$IG(Ex,a) = H(Ex) - \sum_{v \in values(a)} \left( \frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} \cdot H(\{x \in Ex | value(x,a) = v\}) \right)$$

The information gain is equal to the total entropy for an attribute if for each of the attribute values a unique classification can be made for the result attribute. In this case the relative entropies subtracted from the total entropy are 0.

### 3.6 Gain ratio attribute

In this technique we evaluate the worth of an attribute by measuring the gain ratio with respect to the class. By adopting this technique we are able to discriminate against those attributes which have large numbers of distinct values. This fact benefits the usage of gain ratio for many decision-tree based classifiers for improving their accuracies. The attributes with the highest gain ratio are considered most favourable to be used in classification techniques such as the J48, random forest, K-nearest neighbour, naive bayes and SVM, respectively.

The intrinsic value for a test is defined as follows:

$$IV(Ex,a) = -\sum_{v \in values(a)} \frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} * \log_2 \left( \frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} \right)$$

The information gain ratio is just the ratio between the information gain and the intrinsic value:

$$IGR(Ex,a) = IG/IV$$

### 3.7 Symmetrical uncertainty evaluation

In this technique the worth of an attribute is evaluated by measuring the symmetrical uncertainty with respect to the class. We can calculate the symmetric uncertainty of an attribute by SymmU(Class, Attribute) = 2 * (H(Class) - H(Class | Attribute)) / H(Class) + H(Attribute). Here H(Class) & H(Attribute) are entropies for a particular class and a particular attribute**.** Based upon a ranking of the symmetrical uncertainty of the attributes the classification techniques decide which attributes should be given preference over the rest.

## IV. RESULTS

We obtain our search results by using our classification techniques with all the three attribute selection techniques. For all the classifiers the accuracy was over 90 per cent. However two interesting things were observed in the case of random forest method. When the number of attributes and the total number of trees were increasing the accuracy got slightly reduced and in the case of K-nearest neighbour, the larger the neighbourhood the less the accuracy of the classifier. We use two types of data validation which are Cross validation (10 folds) and 50:50 percentage split between testing and the training data sets.

The charts also show us that the accuracy tends to fluctuate with regards to the number of neighbours in the K-neighbourhood algorithm. When we start with only just 1 neighbour then the accuracy is at the peak however slowly the accuracy goes down as we increase the number of neighbours. However the accuracy becomes stagnant at 250 neighbours when we select total number of neighbours at 250 and remains the same till we select the total of neighbours around 1000.

### 4.1 Attribute selection techniques

***Table 1: The performance of the different classifiers while selecting attributes using the information gain technique***

| | Cross validation (10 folds) | | | | Percentage split (50 percent training data -50 percent test data) | | | |
|---|---|---|---|---|---|---|---|---|
| Classifier | Accuracy | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure |
| J48 | 91.9 % | 0.91 | 0.91 | 0.91 | 88.94% | 0.891 | 0.889 | 0.889 |
| Random forest | 95.74% | 0.959 | 0.957 | 0.957 | 94.7% | 0.946 | 0.945 | 0.945 |
| Naive Bayes | 96.17% | 0.962 | 0.962 | 0.962 | 95.74% | 0.958 | 0.957 | 0.957 |
| K-nearest neighbour | 95.11% | 0.953 | 0.951 | 0.951 | 93.19% | 0.937 | 0.932 | 0.932 |
| Support Vector Machine | 95.74% | 0.957 | 0.957 | 0.957 | 95.74% | 0.957 | 0.957 | 0.957 |

***Table 2: The performance of the different classifiers while selecting attributes using the gain ratio technique***

| | Cross validation (10 folds) | | | | Percentage split (50 percent training data -50 percent test data) | | | |
|---|---|---|---|---|---|---|---|---|
| Classifier | Accuracy | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure |
| J48 | 91.9 % | 0.91 | 0.91 | 0.91 | 88.94 % | 0.891 | 0.899 | 0.899 |
| Random forest | 95.74% | 0.959 | 0.957 | 0.957 | 94.47% | 0.946 | 0.945 | 0.945 |
| Naive Bayes | 96.17% | 0.962 | 0.962 | 0.962 | 95.74% | 0.958 | 0.957 | 0.957 |
| K-nearest neighbour | 95.11% | 0.953 | 0.951 | 0.951 | 93.19% | 0.937 | 0.932 | 0.932 |
| Support Vector Machine | 95.74% | 0.957 | 0.957 | 0.957 | 95.74% | 0.957 | 0.957 | 0.957 |

***Table 3: The performance of the different classifiers while selecting attributes using the Symmetrical uncertainty evaluation technique***

| | Cross validation (10 folds) | | | | Percentage split (50 percent training data -50 percent test data) | | | |
|---|---|---|---|---|---|---|---|---|
| Classifier | Accuracy | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure |
| J48 | 91.9 % | 0.91 | 0.91 | 0.91 | 88.94% | 0.891 | 0.889 | 0.889 |
| Random forest | 95.74% | 0.959 | 0.957 | 0.957 | 94.47% | 0.946 | 0.945 | 0.945 |
| Naive Bayes | 96.17% | 0.962 | 0.962 | 0.962 | 95.74% | 0.958 | 0.957 | 0.957 |
| K-nearest neighbour | 95.11% | 0.953 | 0.951 | 0.951 | 93.19% | 0.937 | 0.932 | 0.932 |
| Support Vector Machine | 95.74% | 0.957 | 0.957 | 0.957 | 95.74% | 0.957 | 0.957 | 0.957 |

It is inferable from observing the results from the tables that with our existing data all the classifiers produce the same quality of performance irrespective of attribute selection criteria and entity selection procedure.

The following charts also show us that the accuracy tends to fluctuate with regards to the number of neighbours while using the K-neighbourhood classifier. When we start with only just 1 neighbour then the accuracy is at the peak however slowly the accuracy goes down as we increase the number of neighbours. However the accuracy becomes stagnant at 250 neighbours when we select total number of neighbours at 250 and remains the same till we select the total of neighbours around 1000. We used 50 percent of our data set as training and the rest of the data set as test data.

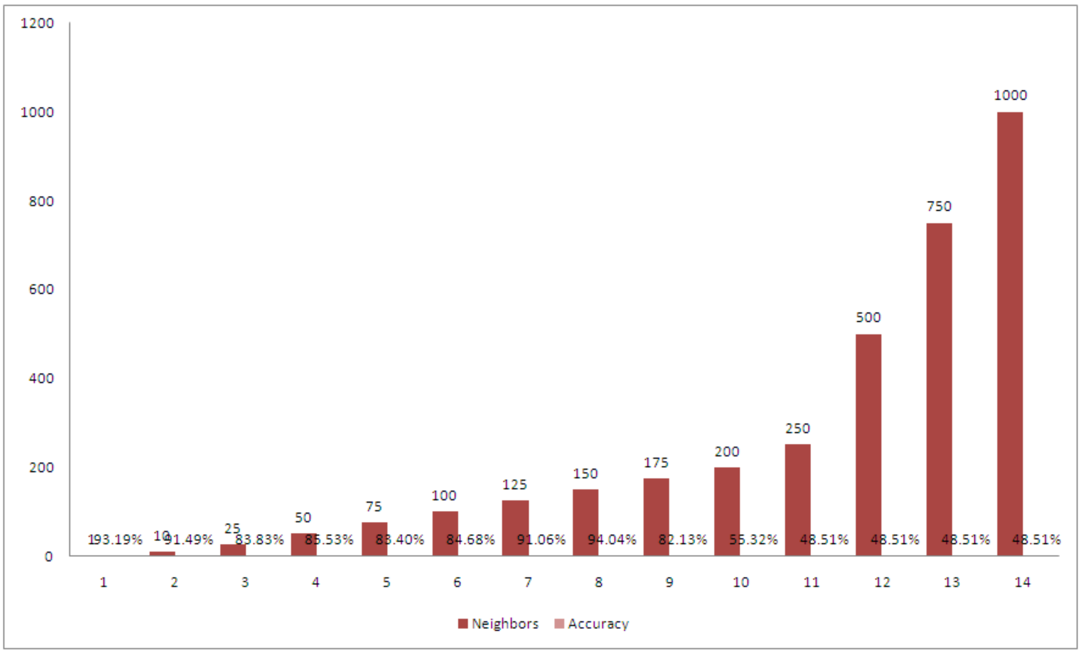

***Figure 2: The impact on accuracy by increasing the number of neighbors in the K-nearest neighbor algorithm***

The random forest technique also behaves similarly to that of the K-neighborhood technique. The accuracy initially dips a little below with the increase in the total number of trees and features and the accuracy figure stabilizes around 91 percent mark with the total number of trees and features both near 1000.

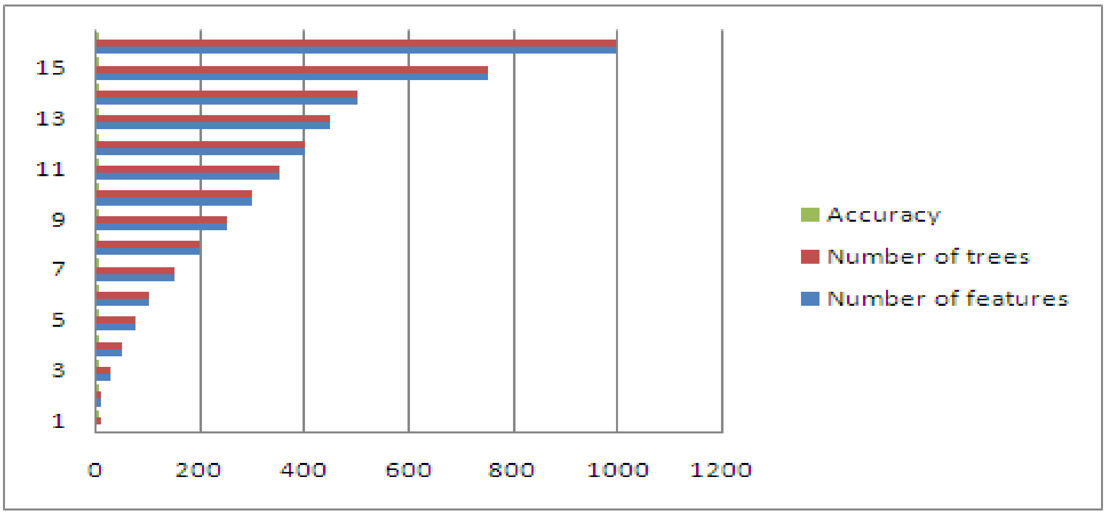

***Figure 3: The impact on accuracy by increasing the number of trees in the random forest algorithm***

## V. CONCLUSIONS

The present work has aimed at creating a diligent technique for categorizing survey answers as either intelligent or egoistic. The attribute selection for the machine learning technique involved three distinct methods i.e. information gain, gain ratio attribute and symmetrical uncertainty evaluation. There are two entity selection procedures i.e. cross validation and 50-50 percentage split which have been adopted under all the attribute selection procedures. The classification techniques which the work deploys are support vector machine, K-nearest neighbour, random forest and the naive bayes method. For the measurement of preciseness of the classification techniques accuracy four parameters i.e. accuracy, precision, recall and f-measure are utilized. The results clearly depict that the performance of all the classification techniques are similar irrespective of the attribute selection process and it is the support vector machine method which performs the best compared to other techniques.

The framework can be useful in presenting a structured technique for indicating how the survey responses are represented. This will enable us to explore the possible correlation between information representation in the surveys and subsequent real world influence of survey opinions. For example the correlation between survey responses and purchasing of memorabilia at online shops could be explored with the help of this framework.

## REFERENCES

I. Wiebe J, Riloff E, Creating subjective and objective sentence classifiers for unannotated texts, *Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, February *13-19*, 2005, 475.

II. Wiebe J, Riloff E, Wilson T Learning subjective nouns using extraction pattern bootstrapping, *In Proceedings of the Seventh Conference on Natual Language Learning, May 31-June 1 , Edmonton , Canada*.

III. Wiebe J, Riloff E, Learning Extraction patterns for subjective expressions *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, 2003, 347-368.

IV. Cambridge dictionary of Psychology, 259.

V. Pang B, Lee L, Opinion mining and sentiment analysis*, Foundations and Trends in Information. Retrieval, 02(01-02)*, 2008, 1-135.

VI. Riloff, E., Wiebe, J., Wilson, T.: Learning Subjective Nouns Using Extraction Pattern Bootstrapping. *In Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003). (2003)* 25–32.

VII. Dave, K., Lawrence, S., Pennock, D.M.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Produce Reviews. *In Proceedings of the 12th International World Wide Web Conference (WWW2003). (2003) Web Proceedings*.

VIII. Wiebe, J., Wilson, T., Bell, M.: Identifying Collocations for Recognizing Opinions. *In Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation. (2001)* 24–31.

IX. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*. (2002) 79–86.

X. Yu, H., Hatzivassiloglou, V.: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003). (2003)* 129–136.

XI. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2000). (2002)* 417–424.

XII. Spertus, E.: Smokey: Automatic Recognition of Hostile Messages. *In Proceedings of the Eighth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-97). (1997)* 1058–1065.

XIII. Das, S.R., Chen, M.Y. Yahoo! for Amazon: Opinion Extraction from Small Talk on the Web. *In Proceedings of the 8th Asia Pacific Finance Association Annual Conference. (2001).*

XIV. Karlgren, J., Cutting, D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *In Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94). (1994)* 1071–1075.

XV. Kessler, B., Nunberg, G., Schütze, H.: Automatic Detection of Text Genre. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97). (1997)* 32–38.