Sentiment Analysis in Student Experiences of Learning

Sunghwan Mac Kim and Rafael A. Calvo
skim1871@uni.sydney.edu.au, rafa@ee.usyd.edu.au
School of Electrical and Information Engineering, University of Sydney

Abstract. In this paper we present an evaluation of new techniques for automatically detecting sentiment polarity (*Positive* or *Negative*) in the students responses to Unit of Study Evaluations (USE). The study compares categorical model and dimensional model making use of five emotion categories: *Anger*, *Fear*, *Joy*, *Sadness*, and *Surprise*. *Joy* and *Surprise* are taken as a *Positive* polarity, whereas *Anger*, *Fear* and *Sadness* belong to *Negative* polarity in the binary classes, respectively. We evaluate the performances of category-based and dimension-based emotion prediction models on the 2,940 textual responses. In the former model, WordNet-Affect is used as a linguistic lexical resource and two dimensionality reduction techniques are evaluated: Latent Semantic Analysis (LSA) and Non-negative Matrix Factorization (NMF). In the latter model, ANEW (Affective Norm for English Words), a normative database with affective terms, is employed. Despite using generic emotion categories and no syntactical analysis, NMF-based categorical model and dimensional model result in better performances above the baseline.

1 Introduction

Universities are increasingly interested in using quality measures that provide evidence that can be used for benchmarking and funding decisions. For this reason, questionnaires such as the Unit of Study Evaluation (USE) [1], or Students Evaluations of Teaching (SET) as they are called in the USA, have been developed as a means of collecting data from students on their experience of learning at the individual subject or unit of study level (these terms are interchangeable and used as synonymously in this study)

Reviews of the literature of student evaluations of teaching show the massive amount of evidence collected using these standard instruments [2, 3]. According to Marsh [2], SET are the most studied form of personnel evaluation. Most of the studies look into quantitative measures gathered in the questionnaires. This paper utilizes the textual open-ended responses that are also collected. Scholarly literature generally agrees on the validity, reliability, dimensionality and actual usefulness of this kind of data. Despite this standardized evaluations have been highly controversial for decades. Arguably because University faculty have normally no formal training in teaching, so those mechanisms that are used for assessing teaching effectiveness are threatening. Staffs are not generally aware of the above mentioned literature, or when they are, they are often skeptic about its meaningfulness.

Despite its criticisms USE and SET are increasingly used by academics, institutions and the students themselves. They reflect valuable aspects of the student experience that can complement other forms of feedback from students to academics and institutions. One of the obstacles is that reading and making sense of all the textual responses can be a daunting task. This paper aims at a combined analysis of the textual and quantitative

responses using novel data mining techniques in order to provide a more comprehensive understanding of the student experience.

Sentiment analysis [4] attempts to automatically identify and recognize opinions and emotions in text. One goal of sentiment classification is to determine whether a text is *objective* or *subjective*, or represents a *positive* or *negative* opinion, affect classification is to identify the expressions of emotion such as *happiness*, *sadness*, *anger*, etc. In the area of sentiment classification, most research has been built around corpora of users' reviews (e.g. movie reviews) that contain a rating system (e.g. number of stars) and a textual description, a data structure similar to the one in the USE used in this study. These reviews are also subjective and contain information about the user experience of the product.

This paper contributes a novel approach to study USE and potentially other descriptions of students' experience. It analyses a corpus of 909 student questionnaires containing 3,353 (raw) textual responses, and explores the feasibility of automatic approaches to making sense of this data. The combination of text and ratings invites the use of sentiment analysis techniques focusing on the valence of opinions (*positive* vs. *negative*), but a richer exploration of student experience would include the affective aspects of the experience beyond its valence. In fact, as it is shown in our study the rating and textual descriptions do not always coincide, possibly highlighting factors in the experience that only surface from the text. The paper explores techniques that could be used for this richer analysis.

The first goal of this paper is to evaluate the feasibility of using sentiment analysis to study the textual responses in USE, an aspect of the data normally sidelined by the ratings. The second goal is to evaluate the merits of two conceptualizations of emotions (*categorical model* and *dimensional model*) on this data.

The rest of the paper is organized as follows: Section 2 presents representative research of the two emotion models used to capture the sentiment of a text. We also describe the affect classifications utilizing the linguistic lexical resources. In Section 3 we will go over the sentiment dataset which comes from USEs. Section 4 provides comparison results from experiments, before coming to our discussion in Section 5.

2 Background

2.1 Emotion Models

There are largely two models for representing emotions: the *categorical model* and *dimensional model*. The categorical model assumes that there are distinct emotional categories such as Ekman's six basic emotions -*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*- [5]. These have been used in many studies despite not necessarily appearing in specific practical scenarios like in teaching. The advantage of such representation is that it represents human emotions intuitively with easy to understand emotion labels. ITS researchers have used other than Ekman's categories to group emotions that appear in

student-system dialogues. D'Mello [6] proposed five categories (boredom, confusion, delight, flow, and frustration) for describing the affect states in ITS interactions.

A second approach where core affects can be represented in a dimensional form [7] represents emotions in a 2 or 3 dimensional space. A valence dimension indicates *positive* and *negative* emotions on different ends of the scale. The arousal dimension differentiates *excited* vs. *calm* states. Sometimes a third, dominance dimension is used to differentiate if the subject feels in control of the situation or not.

The categorical model and the dimensional model have two different methods for estimating the actual emotional states of a person. In the former, a person is required to choose one emotion out of an emotion set that represents the best feeling. On the other hand, the latter exploits rating scales for each dimension like Self Assessment Manikin (SAM) [8], which consists of pictures of manikins, to estimate the degree of valence, arousal, and dominance.

2.2 Categorical Classification with WordNet-Affect

WordNet-Affect [9] is an affective lexical repository of words referring to emotional states. WordNet-Affect extends WordNet by assigning a variety of affect labels to a subset of synsets representing affective concepts in WordNet. In addition, WordNet-Affect has an additional hierarchy of affective domain labels. There are publicly available lists relevant to the six basic emotion categories extracted from WordNet-Affect and we used five lists of emotional words among them for our experiment.

In addition to WordNet-Affect, we exploited Vector Space Model (VSM) in which textual documents can be represented through term-by-document matrix. In general, both terms and documents are encoded as vectors in the reduced *k*-dimensional space. We take into consideration log-entropy with respect to a *tf-idf* weighting schema.

The vector-based representation enables all the contextual information such as words, sentences, and synsets to be represented in a unifying way with vectors. VSM provides a variety of similarity mechanisms between two vectors. In particular, we take advantage of cosine angle between an input vector (input sentence) and an emotional vector (emotional synsets) as similarity measures to identify which emotion the sentence connotes. This can be done in the reduced LSA or NMF representation. We also entail the predetermined threshold ($t = 0.65$) for the purpose of validating a strong emotional analogy between two vectors [10]. If the cosine similarity does not exceed the threshold, the input sentence is labeled as "*neutral*", the absence of emotion. Otherwise, it is labeled with one emotion associated with the closest emotional vector having the highest similarity value. If we define the similarity between a given input text, I, and emotional class, E_j, as $\mathrm{sim}(I, E_j)$, the categorical classification result is more formally represented as follows:

$$\mathrm{CCR}(I) = \begin{cases} \arg\max_j \left(\mathrm{sim}(I, E_j)\right) & \text{if } \mathrm{sim}(I, E_j) \geq t \\ \text{"}neutral\text{"} & \text{if } \mathrm{sim}(I, E_j) < t \end{cases}$$

One class with the maximum index is selected as the final emotion class.

Two statistical dimensionality reduction methods (LSA and NMF) are utilized in a category-based emotion model for the purpose of reducing dimensions in VSM. Graesser and colleagues [11] used Latent Semantic Analysis (LSA) for detecting utterance types and affect in students' dialogue within Autotutor.

Latent Semantic Analysis (LSA) [12] is the earliest model that has been successfully applied to various text manipulation areas. The main idea of LSA is to map terms or sentences into a vector space of reduced dimensionality that is the latent semantic space. The mapping of the given term/sentence vectors to this space is based on singular vector decomposition (SVD). It is known that SVD is a reliable tool available for matrix decomposition. It can decompose a matrix as the product of three matrices. The columns of one of three matrices represent the coordinates for documents in the latent space. Therefore, we make use of the columns in order to compute sentence similarities.

Non-negative Matrix Factorization (NMF) [13] has been successfully applied to semantic analysis. Given a non-negative matrix *A*, NMF finds non-negative factors *W* and *H* that are reduced-dimensional matrices. The product *WH* can be regarded as a compressed form of the data in *A*. This non-negative peculiarity is desirable for handling text data that always require non-negativity constraints. The classification of sentences is performed based on the columns of matrix *H* that represent the sentences.

2.3 Dimensional Estimation with ANEW

ANEW [14] is a set of normative emotional ratings for collections of words (1,035 words) in English, which means that it provides emotional dimensions. This collection provides the rated values for valence, arousal, and dominance for each word that are rated by means of the Self Assessment Manikin (SAM). For each word *w*, the normative database provides coordinates $\overline{w}$ in an affective space as:

$$\overline{w} = (valence, arousal, dominance) = ANEW(w)$$

Therefore, it is possible to accomplish the mapping of contextual information into the 3-dimensional emotion space through ANEW dictionary. For example, words or sentences are scattered all over the emotional plane.

As a counterpart to the categorical classification above, this approach assumes that an input sentence pertains to an emotion based on the least distance between each other on the Valence-Arousal-Dominance (VAD) space. The input sentence consists of a number of words and the VAD value of this sentence is computed by averaging the VAD values of the words. A series of synonyms from WordNet-Affect are used in order to calculate the position of each emotion. These emotional synsets are converted to the 3-dimensional VAD space and averaged for the purpose of producing a single point for the target emotion.

$$\overline{sentence} = \frac{\sum_{i=1}^{n} \overline{w}}{n}, \overline{emotion} = \frac{\sum_{i=1}^{k} \overline{w}}{k}$$

where $\overline{w}$ is the VAD value of a word, and *n* and *k* denote the total number of words in an input sentence and synonyms in an emotion, respectively. *Anger*, *fear*, *joy*, and *sadness* emotions are mapped on the VAD space. Let $Anger_c$, $Fear_c$, Joy_c, Sad_c, and $Surprise_c$ be the centroids of five emotions. Then the centroids, which are calculated by the above equation, are as follows: $Anger_c$ = (2.55, 6.60, 5.05), $Fear_c$ = (3.20, 5.92, 3.60), Joy_c = (7.40, 5.73, 6.20), Sad_c = (3.15, 4.56, 4.00), and $Surprise_c$ = (5.23, 5.33, 4.70). Apart from the five emotions, we manually define *neutral* to be (5, 5, 5). If the centroid of an input sentence ($\overline{sentence}$) is the most approximate to that of an emotion ($\overline{emotion}$), the sentence is tagged as the emotion. We define the distance threshold (empirically set to 4) to validate the appropriate proximity like categorical classification.

3 Unit of Study Evaluation (USE) Data

The Unit of Study Evaluation (USE) questionnaire has 12 questions, 8 of which are standardized University-wide and 4 that are selected by each Faculty. It is designed to provide information to those seeking a) to assess the learning effectiveness of a subject, for planning and implementing changes in the learning and teaching environments, and b) to assess the contributions of units or subjects to students' learning experience in their whole degree program, as monitored by the CEQ. The USE in our study contains 12 statements:

1. The learning outcomes and expected standards of this unit of study were clear to me.
2. The teaching in this unit of study helped me to learn effectively.
3. This unit of study helped me develop valuable graduate attributes.
4. The workload in this unit of study was too high.
5. The assessment in this unit of study allowed me to demonstrate what I had understood.
6. I can see the relevance of this unit of study to my degree.
7. It was clear to me that the staff in this unit of study were responsive to student feedback.
8. My prior learning adequately prepared me to do this unit of study.
9. The learning and teaching interaction helped me to learn in this unit of study.
10. My learning of this unit of study was supported by the faculty infrastructure.
11. I could understand the teaching staff clearly when they explained.
12. Overall I was satisfied with the quality of this unit of study.

Eleven items (I1-I11) focus on students' experience and one item (I12) on student satisfaction. Students indicate the extent of their agreement with each statement based on a 5 - point Likert scale: 1 - strongly disagree, 2 - disagree, 3 - neutral, 4 - agree and 5 - strongly agree. Below each statement there is a space requesting students to explain their response. Question 4 has a different sentiment structure therefore was removed in this study.

The USEs of subjects taught by two academics collected over a period of six years were used to create the dataset. After removing responses to question 4, the dataset contains a total of 909 questionnaires (each with 11 ratings), and out of the possible 9,999, students responded with 3,008 textual responses (each expected to be a description of a rating), a textual response rate of 30.1 %. Out of these we removed internal referencing (e.g. 'see above') and meaningless text (e.g. '?').

The textual data has two characteristics that may significantly affect the classifiers. First the sentences are hand-written in an informal style, containing spelling errors, abbreviated non-dictionary words or hard to read text. The lack of proper grammar would make it extremely challenging to use part-of-speech (POS) tagging or other computational linguistic approaches. Examples include: "Computers in labs too slowk no lecture notes" (spelling mistakes and non-grammar), "tutes were overcrowded, stopping teacher / student interaction" (non-standard words). For these reasons, the techniques used in the experiment are based on the bag-of-words assumption (so word order is not used) and we do not use POS tagging that would require relatively correct grammar.

Table 1. Number of comments and sample comments for each sentiment

Rating	Number	Sentiment	Number	Comments tagged with each sentiment
Strongly Agree	381	Positive	1,455	lecturer and tutor was helpful and explained concepts well.
Agree	1,074			
Neutral	611	Neutral	611	It is a bit clear about staff response but need more examples in there answer.
Disagree	571	Negative	874	Not enough computers to accommodate all the students.
Strongly Disagree	303			

4 Experiments and Results

The following five different approaches are implemented in Matlab. One categorical model that has two variants, according to three corresponding methods of dimension reduction, one dimensional method, and two similarity comparison methods for each model are implemented. For evaluation purposes, we employ Majority Class Baseline (MCB) as our baseline and Keyword Spotting (KWS). We remove stop words and use stemming. Text to Matrix Generator (TMG), a Matlab toolkit [15], is used to generate term-by-sentence Matrix.

- Majority Class Baseline (MCB): classification that always predicts the majority class, which in this dataset is *Positive* across all sentiment classifications.
- Keyword Spotting (KWS): a naïve approach that counts the presence of obvious affect words like "frustrating" and "satisfaction", which are extracted from WordNet-Affect for five emotion categories.
- CLSA: LSA-based categorical classification
- CNMF: NMF-based categorical classification
- DIM: Dimension-based estimation

Five emotion categories are utilized (*Anger*, *Fear*, *Joy*, *Sadness*, and *Surprise*) in which *Joy* and *Surprise* emotions are assigned to *positive* class while *Anger*, *Fear*, and *Sadness* are the members of *negative* class, respectively. Negative emotion, *disgust*, is removed because the emotion is similar to *anger* and leads to making sentiment classes biased. Likewise, *strongly agree* and *agree* belong to *positive*, and *strongly disagree* and *disagree* are referred to *negative*. The number of sentences for each rating and sentiment used in our experiment is shown in Table 1. In addition, sample comments of the annotated corpus appear in Table 1.

Table 2 shows the precision, recall, and F-measure values obtained by the five approaches for the automatic classification of three sentiments. The highest results are marked in bold for each individual class. We do not include accuracy values in our results due to the imbalanced categories (see Table 1). The accuracy metric does not provide adequate information, whereas precision, recall, and F-measure can effectively evaluate the classification performance with respect to imbalanced datasets [16].

Table 2. Sentiment identification results

Sentiment	Positive			Negative			Neutral		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
MCB	0.495	**1.000**	**0.662**	0.000	0.000	-	0.000	0.000	-
KWS	0.527	0.220	0.310	0.270	0.061	0.099	0.212	**0.743**	**0.330**
CLSA	0.575	0.362	0.445	0.388	0.203	0.266	0.218	0.560	0.314
CNMF	0.505	0.897	0.646	0.378	0.120	0.182	**0.421**	0.052	0.093
DIM	**0.591**	0.329	0.423	**0.398**	**0.317**	**0.353**	0.223	0.522	0.312

As can be seen from the table, the performances of each approach depend on each sentiment category. In case of the *positive* class, which has the largest number of sentences, MCB and CNMF get the best sentiment detection performance in terms of recall and F-measure. DIM achieves rather high precision score in comparison with all other classifications. We can see that DIM approach gives the best results for *negative* class. When it comes to *neutral*, KWS shows the best performance with respect to recall and F-measure. On the other hand, CNMF particularly outperforms the others for precision. Figure 1 indicates a result of the 3-dimensional and 2-dimensional attribute evaluation for USEs.

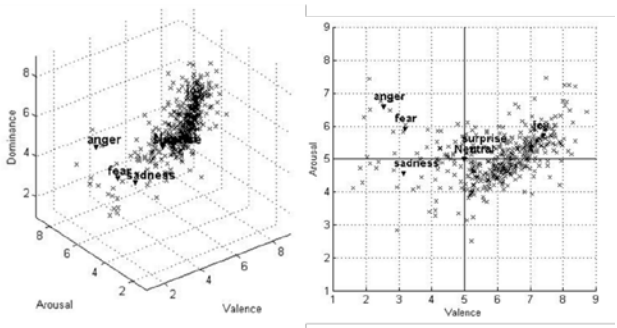

Figure 1. Distribution of the USEs dataset in the 3-dimensional (left) and 2-dimensional (right) sentiment space. The 'x' denotes the location of one comment corresponding to valence, arousal, and dominance.

A notable aspect observed in the USE data is that there are somewhat inconsistencies between students' ratings and written responses illustrated with examples in Table 3. For instance, the third row is unambiguously negative but the student graded this sentence as neutral. Therefore, all approaches have a weakness in recognizing sentiments due to the peculiarity of this data. Another factor, which makes the automatic classification difficult, is that all classifiers are not specific to education domains. For this reason, we speculate that this mediocre performance of the methods is owing to poor coverage of the features found in education domains.

Table 3. Sample feedbacks from misclassified results. (Positive values are those rates 4 as 5, neutral as 3 and negative 1 or 2)

Student's feedback	Student rating	System rating
It should be core to software gingerbeering	Positive (5)	Neutral (LSA)
The labs were not long enough with too few tutorials. 4 labs were too few. How about one for FETS/MOSFETS? Given the instruction was for AC/DC components (i.e. lower/uppercase) it was difficult to follow the hadn written notes on the overhead. Maybe print it all up?	Positive (4)	Negative (NMF)
We never got personal feedback.	Neutral (3)	Negative (DIM)
Hi my name is ABC, I like this LECTURER_NAME, I mean this course!!	Negative (2)	Positive (NMF)

Table 4 shows overall precision, recall, and F-measure comparison with respect to MCB, KWS, CLSA, CNMF, and DIM in two averaging perspectives (micro-averaging and macro-averaging). The notable difference between these to calculate is that micro-averaging gives equal weight to every sentence whereas macro-averaging weights equally all the categories. From this summarized table, we can see that MCB, KWS, and CLSA perform less effectively with a little low number of evaluation scores compared with CNMF and DIM. In case of macro-averaging, CNMF is superior to other classifications in precision, while DIM surpasses the others in recall and F-measure. On the other hand, DIM has the best precision and CNMF performs better for F-measure in micro averaging. Overall, CNMF and DIM vie with each other in precision, recall and F-measure and the best F-measure is obtained with the approach based on CNMF or DIM for each average. Our KWS conducted in all experiments is inferior to CNMF, DIM as well as CLSA. The result implies that keyword spotting techniques cannot handle the sentences which evoke strong emotions through underlying meaning rather than affect keywords. In addition, we can infer that the models (CNMF and DIM) with non-negative factors are appropriate for dealing with text collections. In summary, NMF-based categorical model and dimensional model shows the better sentiment recognition performance as a whole.

The most frequent words used by students to describe aspects of their experience, include terms such as *labs*, *lecturer*, *lectures*, *students*, *tutors*, *subject*, and *work*. When we remove these terms, the words most frequently used to describe positive experiences include: *good* (*n*=263), *helpful* and *helped* (*n*=183), *online* (*n*=79), *understand* (*n*=49). Those used to describe negative experiences include: *hard* (*n*=72), *understand* (*n*=67),

time (*n*=47). Neutral experiences contain a combination of both. These words lists are obtained from CNMF and DIM because two classifications have better overall performance as aforementioned. Stemming was not used for this analysis since in this particular corpus it might hide important differences as between ‘lecturer’ and ‘lecture’.

Table 4. Overall average resutls

Mean	Micro			Macro		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
MCB	0.245	**0.495**	0.328	0.165	0.333	0.221
KWS	0.385	0.281	0.325	0.337	0.341	0.247
CLSA	0.445	0.356	0.396	0.394	0.375	0.342
CNMF	0.450	0.490	**0.469**	**0.434**	0.356	0.307
DIM	**0.457**	0.366	0.406	0.404	**0.389**	**0.363**

5 Discussion

This paper described a dataset of ratings and textual responses of student evaluations of teaching. Sentiment analysis techniques for automatically rating textual responses as *positive*, *negative* or *neutral* using the students’ ratings were evaluated. In particular, the performance of categorical model and dimensional model were compared, each of which makes use of different linguistic resources.

This paper highlighted that NMF-based categorical and dimensional models have a better performance than the others. Moreover, despite not having an appropriate set of emotional categories to use, the efficacy of two emotion lexicons (WordNet-Affect and ANEW) promises to be useful in these sentiment classification tasks.

While two models and two lexicons are promising for identifying sentiments, there are still challenges to overcome. We believe that affective expressivity of text is on the basis of more complex linguistic features such as morphological features. Hence, we are going to delve into Natural Language Processing (NLP) to recognize fine-grained emotion in the future.

Future work will include extending the corpora with more student evaluations and this should provide more reliable results. The categorical model should be evaluated with a set of emotion categories better grounded in the educational research literature and we suspect that the literature on motivation would be particularly useful. With regards to the use of normative databases to study the dimensional model, we are aware that the terms in ANEW are not the best suited for the vocabulary that students use to describe their experiences, but we are not aware of other more appropriate databases.

Acknowledgement

This project was partially funded by a TIES grant from the University of Sydney.

References

[1] ITL. *Purpose of the USE* 2008 [cited 2009 February 10]; Available from: http://www.itl.usyd.edu.au/use/purpose.htm.

[2] Marsh, H.W., Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 1987. **11**(3): p. 253-388.

[3] Richardson, J.T.E., *Instruments for obtaining student feedback: a review of the literature.* Assessment & Evaluation in Higher Education, 2005. **30**(4): p. 387-415.

[4] Pang, B. and L. Lee, *Opinion mining and sentiment analysis.* Foundations and Trends in Information Retrieval, 2008. **2**(1-2): p. 1-135.

[5] Ekman, P., *An Argument for Basic Emotions*. 1992. p. 200.

[6] D'Mello, S., R. Picard, and A. Graesser, *Toward an Affect-Sensitive AutoTutor.* IEEE Intelligent Systems, 2007. **22**(4): p. 53-61.

[7] Russell, J.A., *Core affect and the psychological construction of emotion.* Psychological Review, 2003. **110**(1): p. [Washington, etc.] American Psychological Association [etc.]--172.

[8] Lang and P.J. (1980). "Behavioral treatment and bio-behavioral assessment: Computer applications." Technology in mental health care delivery systems: 119-137.

[9] Strapparava, C. and R. Mihalcea. Learning to identify emotions in text. in Proceedings of the 2008 ACM symposium on Applied computing. 2008.

[10] Penumatsa, P., M. Ventura, et al. (2006). The Right Threshold Value: What Is the Right Threshold of Cosine Measure When Using Latent Semantic Analysis for Evaluating Student Answers? International Journal on Artificial Intelligence Tools, WORLD SCIENTIFIC PUBLISHING.

[11] D'Mello, S., et al. Automatic Detection of Learner's Affect from Conversational Cues. in User Modeling and User-Adapted Interaction. 2008.

[12] Landauer, T.K., et al., eds. *Handbook of Latent Semantic Analysis*. 2007, Routledge.

[13] Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**(6755): p. 788-791.

[14] Bradley, M. M. and P. J. Lang (1999). "Affective norms for English words (ANEW): Instruction manual and affective ratings." University of Florida: The Center for Research in Psychophysiology.

[15] Zeimpekis, D. and E. Gallopoulos, *TMG: A MATLAB toolbox for generating term-document matrices from text collections.* Grouping multidimensional data: Recent advances in clustering, 2005: p. 187-210.

[16] He, H. and E.A. Garcia, *Learning from Imbalanced Data.* IEEE Transactions on Knowledge and Data Engineering, 2009. **21**(9): p. 1263.