International Conference on Latest Advancements & Future Trends In Engineering, Science & Management
Sambhram Institute of Technology, Bangalore (India)
03rd - 04th May 2019 www.conferenceworld.in
SAMBHRAM INSTITUTE OF TECHNOLOGY
ISBN : 978-93-87793-86-6

Review Summarization and Aspect Category Detection with Co-occurrence data by refining Word Embeddings

Gaurav Sharma[1], Milind Waghmare[2]

1 (Department of Computer Science and Engineering,

2 Government College of Engineering, Amravati, India)

3 (Department of Computer Science and Engineering,

4 Government College of Engineering, Amravati, India)

ABSTRACT

Online customer reviews is becoming crucial in customer decision making. The number of reviews has been increasing and it is quite difficult to analyse all such reviews and form opinions about products or services. Hence, there is a need for a system that can summarize these reviews and present it in concise form to user. The sentiments and context of the reviews should be preserved in such summary. Aspect level sentiment analysis can help in summarizing various aspects of a product or service. The challenge is to detect aspect categories from reviews and then associate suitable sentiments to it. The proposed system do this using Support Vector Machine (SVM). Also, word embedding and corresponding polarity score is used to better capture the sentiments of reviews.

Keywords- Aspect category detection, machine learning, natural language processing. Sentiment analysis, word embedding.

1. Introduction

Electronic reviews has replaced word of mouth (WoM) in consumer decision making process. Reviews are considered more important than editorial recommendations. E-commerce sites have numerous reviews for the products they sell. The information from these reviews is not only beneficial for customers but also for the companies. They can be used to track buyer preference and grievances. This can help companies improve their services or products.

However, the number of reviews are ever-increasing and it is becoming difficult to analyse all these and form opinion. Add it to the fact that there are competing products which have their own set of reviews and it becomes tedious for consumer to analyse and compare. Also, average ratings are not as useful as they fails to convey much information about products or services. Hence, it is necessary to have a model which process the information from reviews and present it to user in summarized form.

One task for such model is to extract sentiments from reviews. However, to preserve fine grained information while extracting summary, aspect level analysis is needed. Thus, the model needs to first determine aspect topics and then its corresponding sentiment. The topics could be very fine grained or generalized depending on the domain and requirement. For example, fine grained aspect topics from hotel reviews are “egg”, ”fish” and “rice”

while the general aspect category is "food".

Also, it is computationally costly to process natural language. Hence, the words are represented as low dimensional real valued vectors. This representation is called Word Embeddings. It helps in capturing syntactical and semantic relationship between words effectively. In such representation, similar words are classified into similar vectors and dissimilar words into dissimilar vectors. Since, it is represented in numeric format, it is easy to process words and their relationship.

2. Related Works

Aspect categories are usually implied as they are not usually explicitly mentioned in the sentences and needs to be inferred. This makes it difficult to detect aspect categories. Early work of implicit category detection is mentioned in [1]. The proposed approach is to make use of semantic association analysis in differentiating implicit aspect categories from notional words. In [2], unsupervised approach is suggested to simultaneously cluster aspect and opinion words. It does that by combining both homogeneous similarity and heterogeneous relationships information between opinion words.

In [3], a two-step co-occurrence rule mining approach is proposed to determine implicit aspects. In first phase, association rule are mined from co-occurrence matrix. The entry in matrix represent frequency of opinion word co-occurring with explicitly mentioned aspect. In next phase, the rule consequents are clustered to obtain more robust rule for every opinion word. By identifying best cluster for sentiment word with no explicit aspect, implicit aspect could be found. In [4], a semi-supervised method simultaneously extract both aspect and sentiment words from review sentences. Association rule mining is also utilized in [5] and [6]. In [7], a high performing supervised method for category detection is presented. In proposed method, binary entropy classifier and TF-IDF methods are used for aspect category detection. Garcia-Pablos et al. [8] proposed unsupervised approach in which first aspects are detected and then compared with category words using similarity index. In [9], a comparison of supervised and unsupervised methods is made. Both method uses co-occurrence association rule mining to detect categories. A detailed study of aspect detection is presented in [10].

Word embeddings uses learning methods for word representation in real valued vector form. Word embeddings generated from particular context tend to generate similar representation for contextually similar words. This works well for semantically similar words but sentimentally opposite words also tend to be represented by similar vectors. This problem is mentioned in [11] and [12]. Sentiment embeddings have been proposed to avoid similar vector generation for sentimentally opposite words in [13] and [14]. It captures both semantic and sentiment information so that sentimentally similar words have similar vector representation. In [15] and [16], convolutional network is used to incorporate preceding and succeeding contexts into word embedding. [17] Used topic and sentiment information in single prototype model for sentiment embeddings and then modified it into multi-prototype model.

In [18], k-Nearest Neighbour is applied on existing word embedding for refining it using intensity score for better sentiment representation.

3. Methodology

The proposed system as shown in fig. 1 processes review dataset and summarize it for better reading. The hotel dataset from the SemEval-2014 Challenge [19] is used in the system. The system consists of three components namely user section, hotel section and admin section. Although, the system is domain specific, it can easily be extended to other generalized domains.

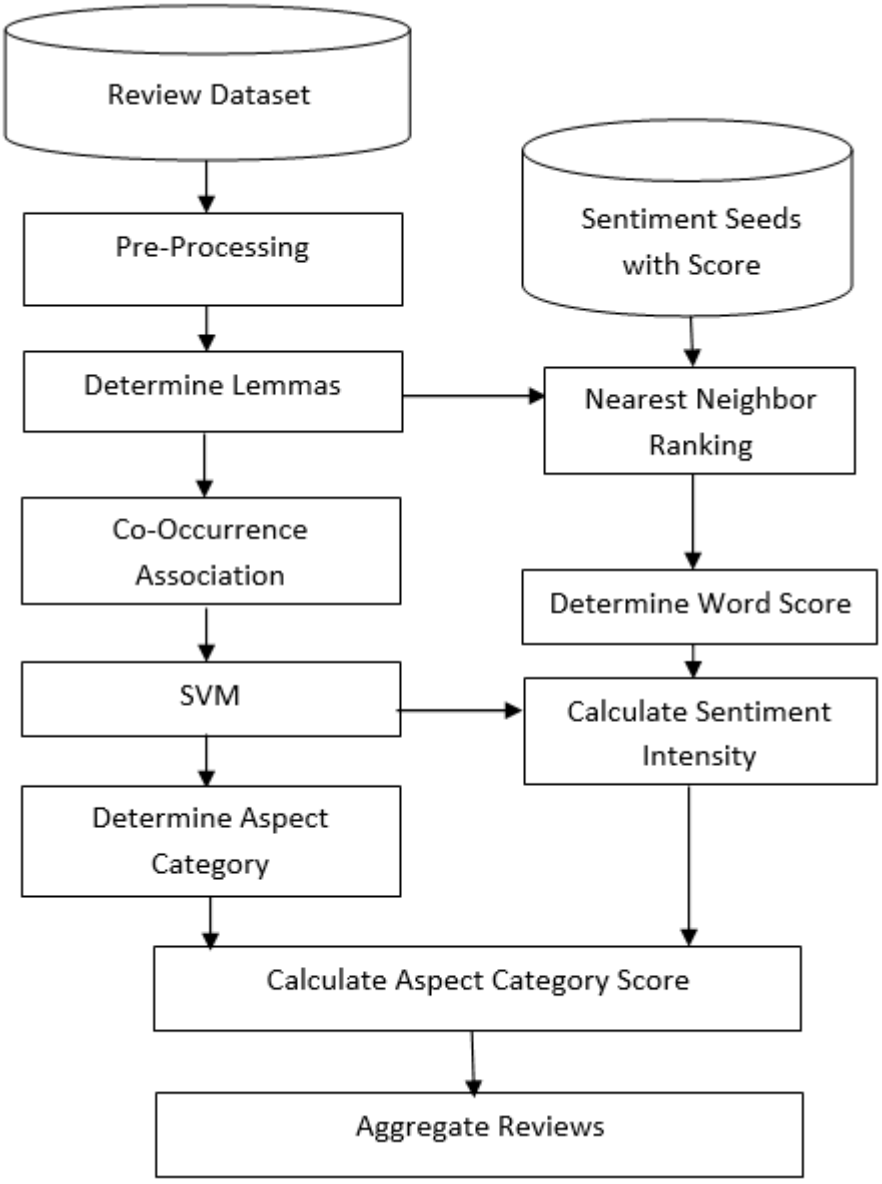


Fig. 1 System Architecture

3.1 User Section

The user can register and login in the system. The user can view the list of hotels, their description and hotel reviews. It can also post its own review on a hotel and give star rating. The average of all star ratings is also calculated and shown on hotel page. On hotel detail page, a summary of reviews based on detected aspect is presented to the user.

3.2 Hotel Section

The hotel owner can also register and login in the system. It can set the details of hotels for the users to see. It can also view all the reviews on its hotel and also it's summarized form.

3.3 Admin Section

The admin can see the list of all registered user and hotel owners. It can also block or allow any user or hotel owner. The admin can also view all hotels and their details. It can trigger the review summarization process for any hotel and then view its review summary.

The review summarization process consists of following stages:

3.3.1 Data pre-processing

The review dataset is first pre-processed. The dataset is cleansed so that further processing can be done effectively. All the white spaces and special characters are removed. Also, the reviews are split into each individual sentences. Then, stop words are removed from each sentences as shown in fig.2. This is done by using pre-determined set of stop words. The Porter algorithm is applied for stemming. This results in all the words in their root form and helps in effectively processing them.

Stop Word Removal

	Processed Review	Stop Words
1	Food was tasty but so expensive.	food tasty expensive
2	To be completely fair, the only redeeming factor was the food, which was above average, but could not make up for all the other deficiencies of Teodora.	completely fair redeeming factor food average not make deficiencies teodora
3	The food is uniformly exceptional, with a very capable kitchen which will proudly whip up whatever you feel like eating, whether its on the menu or not.	food uniformly exceptional capable kitchen will proudly whip whatever feel like eating whether menu not
4	Where Gabriela personaly greets you and recommends you what to eat.	gabriela personaly greets recommends eat
5	For those that go once and do not enjoy it, all I can say is that they just do not get it.	go not enjoy can say just not get
6	Not only was the food outstanding, but the little perks were great.	not food outstanding little perks great
7	It is very overpriced and not very tasty.	overpriced not tasty

Fig. 2. Stop Word Removal

3.3.2 Aspect Detection

Aspect detection is an important step and co-occurrence matrix is computed first for aspect category detection. Stanford parser [20] is used to process and form dependency relations. The co-occurrence frequencies is used to identify candidates for aspect categories. Low frequency dependency are ignored to avoidover-fitting. Then, Support Vector Machine (SVM) is used to detect categories as in fig. 3 since it can handle unbalanced data [21]. After calculating co-occurrence frequencies lemma-category combinations and forming a co-occurrence matrix, associated conditional probability is calculated and stored as weights. Optimal threshold is then determined based on the training dataset for category detection. If the weights allotted to a dependency is greater than the optimal threshold then that aspect category is assigned to that sentence.

Aspect Category Detection

	Aspect Words	Aspect Category
1	food	food
2	factor	service
3	food	food
4	food	food
5	kitchen	food
6	whip	misc
7	feel	service
8	eat	food

Fig. 3. Aspect Category Detection using SVM

3.3.3 Word Embedding and PolarityCalculation

For better sentiment analysis, the words obtained after stemming process are represented as vectors. This is called word embedding. Using nearest neighbour ranking algorithm, this word vectors are then mapped to words from pre-defined dataset. Then, an intensity score is calculated as in fig. 4 for the given word based on the neighbours found in the ranking algorithm. For example, "good" and "great" will have similar intensity score with "great" having more intensity score than "good". This intensity score will not only help in determining the sentiments for detected aspect categories but also help in better conveying the emotions expressed in the review instead of just positive or negative.

Sentiment Words

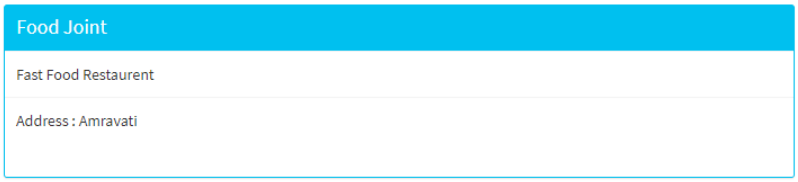
	Sentiments	Polarity
1	tasty expensive	tasty(3.5) expensive(-3.0)
2	fair average deficiencies	fair(2.0) average(2.5) deficiencies(-2.0)
3	exceptional capable proudly like	exceptional(4.5) capable(1.0) proudly(2.0) like(2.0)
4	greets recommends	greets(1.0) recommends(2.0)
5	not-enjoy	not-enjoy(-2.0)
6	outstanding great	outstanding(5.0) great(3.0)
7	overpriced not-tasty	overpriced(-3.5) not-tasty(-3.5)
8	agreed favorite kind	agreed(1.0) favorite(2.0) kind(2.0)
9	outstanding terrific	outstanding(5.0) terrific(4.0)

Fig. 4. Determine Sentiment and its Polarity

3.3.4 Review Aggregation

The sentiment and intensity score are then associated with the detected categories. An average or most prominent intensity score is then identified for each aspect. All the detected aspects and their corresponding associated sentiment word/ intensity score is presented to the user. This can be done either by showing average rating (ex. from 1 to 5) or by showing sentiment words (ex. good, bad, awesome, etc.) corresponding to intensity score of the aspect category. Thus, a summary of each detected aspect is presented to user for analysis and

comparison.



Aggregated Data

	Aspect Category	Score	Star
1	AMBIENCE	6.6	★★★★★
2	PRICE	7.3	★★★★★
3	SERVICE	6.2	★★★★★
4	FOOD	7.2	★★★★★
5	MISC	4.8	★★★★★

Fig. 5. Aggregate Review for each Aspect Category

4. Result

The result of the system is measured by the parameters Recall, Precision, F-Measure and Accuracy. As seen in fig. 6 the system has slight less recall but high precision. Fig. 7 represent the relative percentage of aspect terms detected for each aspect category.

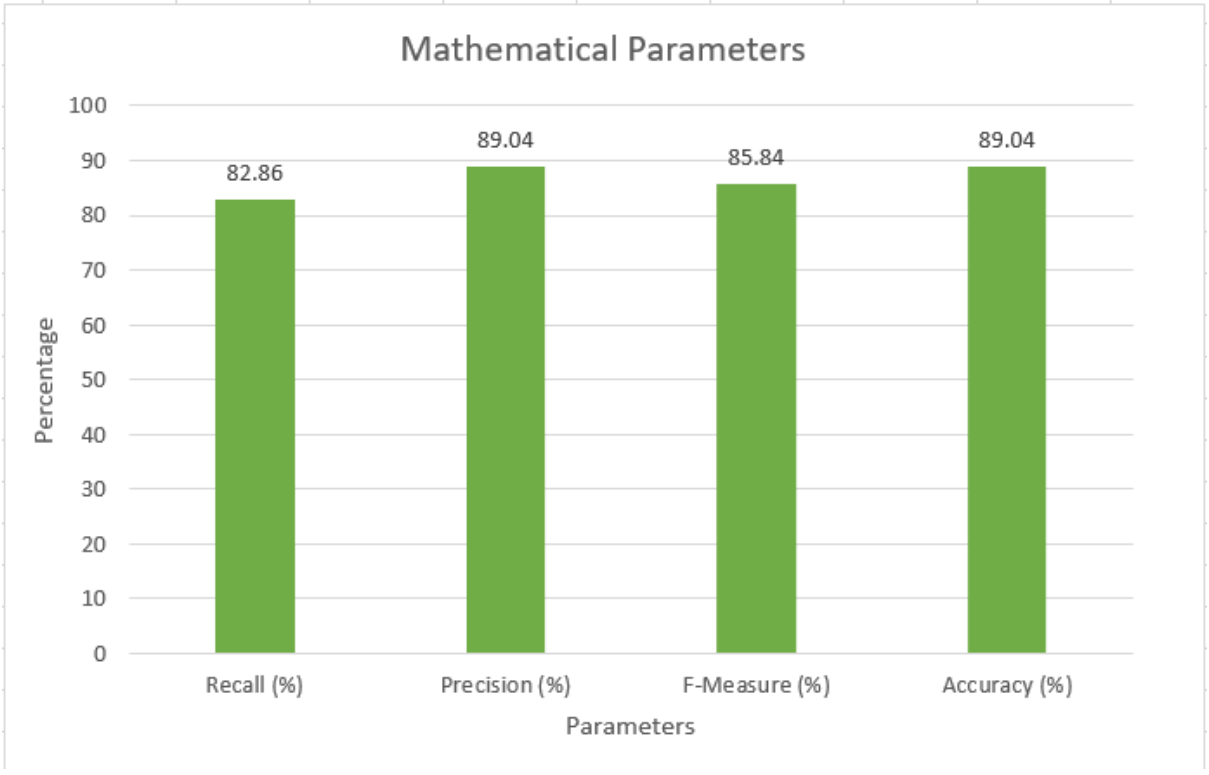


Fig. 6. Mathematical Parameters

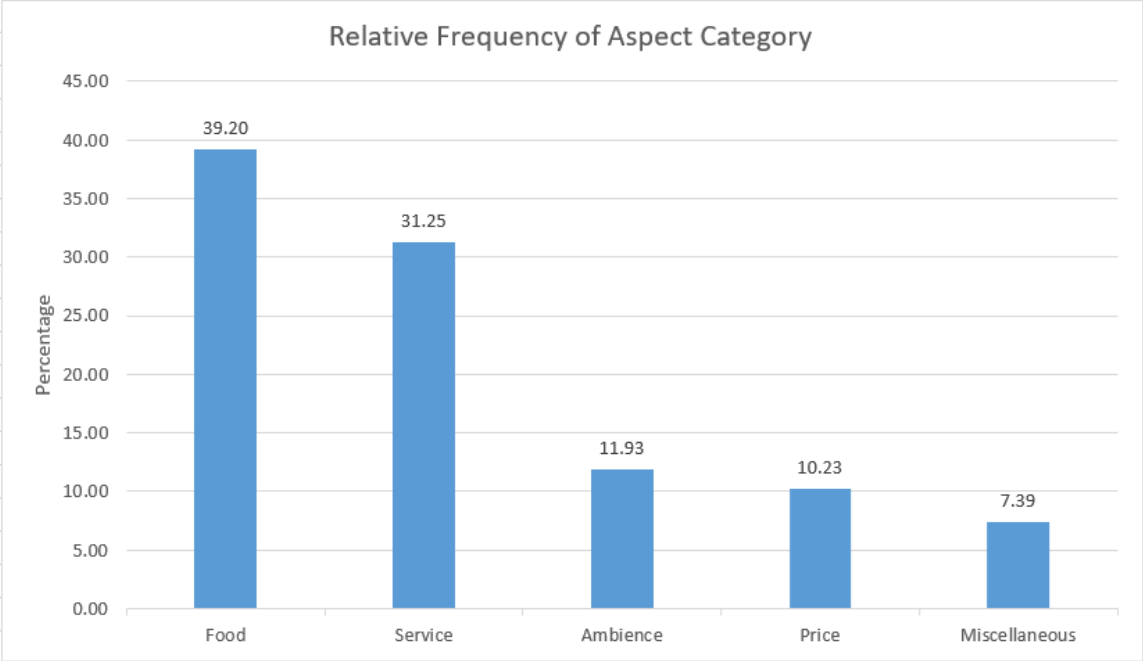


Fig. 7. Relative frequency of detected aspect categories

5. Conclusion

A review summarization system is proposed which detect aspect level sentiments and present the review in summarized form for specified aspect categories of hotels like price, food quality, ambience, etc. Using co-occurrence matrix and SVM, aspect categories are detected. Word embeddings and intensity score is used to better capture the sentiments in reviews. This also helps in better calculating the ratings of each aspect category. The summarized review for each aspect category and its corresponding score helps user in better decision making as comparison can be made according to various aspects of product or services. Although the proposed system is based on a specific domain (i.e. hotel reviews), it can easily be incorporated into another or more generalized domain.

REFERENCES

[1] Q. Su, K. Xiang, H. Wang, B. Sun, and S. Yu, "Using pointwise mutual information to identify implicit features in customer reviews," in Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead (LNCS 4285), Y. Matsumoto, R. Sproat, K.-F. Wong, and M. Zhang, Eds. Berlin, Germany: Springer, 2006, pp. 22–30.

[2] Q. Su et al., "Hidden sentiment association in Chinese Web opinion mining," in Proc. 17th Conf. World Wide Web (WWW), Beijing, China, 2008, pp. 959–968.

[3] Z. Hai, K. Chang, and J.-J. Kim, "Implicit feature identification via co-occurrence association rule mining," in Proc. 12th Int. Conf. Comput. Linguist. Intell. Text Process. (CICLing), Tokyo, Japan, 2011, pp. 393–404.

[4] X. Zheng, Z. Lin, X. Wang, K.-J. Lin, and M. Song, "Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification," Knowl. Based Syst., vol. 61, no. 1, pp. 29–47, 2014.

[5] W. Wang, H. Xu, and W. Wan, "Implicit feature identification via hybrid association rule mining," Expert Syst. Appl. Int. J., vol. 40, no. 9, pp. 3518–3531, 2013.

[6] Y. Zhang and W. Zhu, "Extracting implicit features in Online customer reviews for opinion mining," in Proc. 22nd Int. Conf. World Wide Web Companion (WWW Companion), 2013, pp. 103–104.

[7] T. Brychcin, M. Konkol, and J. Steinberger, "UWB: Machine learning approach to aspect-based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 817–822.

[8] A. Garcia-Pablos, M. Cuadros, S. Gaines, and G. Rigau, "V3: Unsupervised generation of domain aspect terms for aspect based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 833–837.

[9] Kim Schouten, Onne van der Weijde, Flavius Frasincar, and Rommert Dekker, "Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data", IEEE Transactions on Cybernatics, vol. 48, no. 4, 2018.

[10] Kim Schouten and Flavius Frasincar, "Survey on Aspect-Level Sentiment Analysis", IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 3, March 2016.

[11] S. M. Mohammad, B. J. Dorr, G. Hirst, and P. D. Turney, "Computing lexical contrast," Comput. Linguistics, vol. 39, no. 3, pp. 555–590, 2013.

[12] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment embeddings with applications to sentiment analysis," IEEE Trans. Knowl. Data Eng., vol. 28, no. 2, pp. 496–509, Feb. 2016.

[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Ng, and C. Potts, "Learningword vectors for sentiment analysis," in Proc. ACL, 2011, pp. 142–150.

[14] I. Labutov and H. Lipson, "Re-embedding words," in Proc. ACL, 2013, pp. 489–493.

[15] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proc. Int. Conf. Mach. Learn., 2008, pp. 160–167.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.

[17] Y. Ren, Y. Zhang, M. Zhang, and D. Ji, "Improving Twitter sentiment classification using topic-enriched multi-prototype word embeddings," in Proc. AAAI, 2016, pp. 3038–3044.

[18] Liang-Chih Yu, JinWang, K. Robert Lai, and Xuejie Zhang, "Refining Word Embeddings Using Intensity Scores for Sentiment Analysis", IEEE/ACM Transactions on Audio, Speech, and Laguage Processing, VOL. 26, NO. 3, March 2018.

[19] M. Pontiki et al., "SemEval-2014 Task 4: Aspect based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 27–35.

[20] C. D. Manning et al., "The Stanford CoreNLP natural language processing toolkit," in Proc. 52nd Annu. Meeting Assoc. Comput. Linguist. Syst. Demonstrations, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010.

[21] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 39, no. 1, pp. 281–288, Feb. 2009.