iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69980

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Sentiment Analysis Based on Category Detection Using Machine Learning Techniques

Mr. Chittimuru S Reddy[1], Prof. V. V. Sunil Kumar[2]
PBR Visvodaya Institute of Technology & Science, India

Abstract: ***In this paper, the online consumer reviews were considered to assist purchase- decision making has become increasingly popular. To process the user reviews and find the useful information for making decision of purchase most of existing systems are presented. But one can hardly read all reviews to obtain a fair evaluation of a product or service. A subtask to be performed by such a framework would be to find the general aspect categories addressed in review sentences, for which this project presented two methods. The first method presented is an unsupervised method that applies association rule mining on co-occurrence frequency data obtained from a corpus to find these aspect categories. While not on par with state-of-the-art supervised methods, the proposed unsupervised method performs better than several simple baselines, a similar but supervised method, and a supervised baseline, with an F1-score of 67%. The second method is a supervised variant that outperforms existing methods with an F1-score of 84%.***

I. INTRODUCTION

Data mining might even be a term from engineering. Usually it's in addition remarked as information discovery in databases (KDD). Process is relating to finding new data terribly ton of data. The knowledge obtained from process is hopefully each new and helpful. In several cases, data is hold on thus it is typically used later. The info is saved with a goal. As degree example, a store desires to avoid wasting what has been bought. They have to undertake to this to grasp what proportion they have to induce themselves, to own enough to sell later. Saving this data, makes myriad data. the data is typically saved terribly data. The principle why data is saved is named the primary use.

Later, a similar data might even be accustomed get completely different data that wasn't required for the primary use. The search ought to grasp presently what quite things individuals get on once they search the search. (Many those that get food in addition get mushrooms as degree example.) That sort of data is among the knowledge, and is useful; however wasn't the principle why the info was saved. This data is new and might be helpful. It's a second use for a similar data.

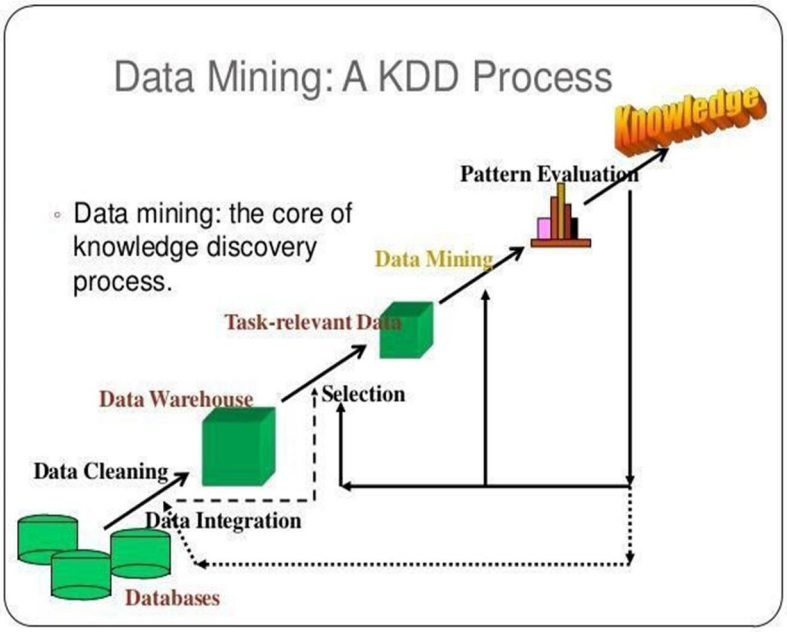


Figure 1.1: KDD Process

A. About Sentiment Analysis

Word of Mouth (WoM) has always been influential on consumer decision- making. Family and friend are usually asked for advice and recommendations before any important purchase-decisions are made. These recommendations can both have short similarly as long stretch impact on buyer dynamic [1].

With the Web, WoM has massively expanded. Any person who wishes to share their experiences would now have the option to do so electronically. Online media, like Twitter and Facebook think about basic ways to deal with exchange enunciations regarding things, organizations, and brands. The articulation for this all-inclusive kind of WoM is electronic WoM (EWoM).

Over the span of the two or three years, EWoM has become dynamically renowned [2]. Maybe the fundamental kinds of EWoM correspondence are thing and organization reviews [3] posted on the Web by clients. Retail associations, for instance, Amazon and Bol have different reviews of the things they sell, which give a plenitude of information, and objections like Yelp offer bare essential buyer overviews of neighborhood bistros, lodgings, and various associations. Investigation has shown these overviews are seen as more critical for customers than market-created information and article ideas [4]–[6], and are logically used in purchase dynamic [7].

The information that can be gotten from thing and organization reviews isn't just helpful to purchasers, yet also to associations. Acknowledging what has been posted on the Web can help associations with improving their things or organizations [8].

In any case, to effectively manage the colossal proportion of information available in these reviews, a framework for the robotized outline of reviews is appealing [9]. A huge endeavor for such a construction is see the subjects (i.e., characteristics of the thing or organization) people clarify. These subjects can be fine-grained, because of viewpoint level speculation examination, or more customary by virtue of perspective classes.

As ought to be self-evident, point characterizations are by and large recommended, that is, the names of the groupings are not unequivocally referred to in the sentence.

Right when the point of view orders are known, and enough planning data is available, a guided AI approach to manage perspective class distinguishing proof is conceivable, yielding an unrivaled [11]. Various approaches to manage find point of view classes are directed [11]–[14]. Nevertheless, to a great extent the versatility characteristic for an independent strategy is alluring.

In this endeavor, both a performance and a controlled strategy are suggested that can find perspective classes reliant upon co-occasion frequencies. The independent system uses spreading sanctioning on an outline worked from word co-occasion frequencies to recognize viewpoint orders. Moreover, no assumption should be made that the suggested points are continually implied unequivocally, like it is done in [15]. The proposed solo method uses something past the demanding arrangement mark by making a lot of unequivocal lexical depictions for each order. The singular required information is the plan of point classes that is used in the instructive record. The controlled method of course uses the co-occasions between words, similarly as phonetic association fundamentally increments, and the remarked on perspective arrangements to discover unforeseen probabilities from which ID rules are mined.

What's more, moreover we loosen up our work to oversee imbalanced data using Synthetic Minority Over-testing Technique (SMOTE). Further improving execution we adjusted SMOTE.

II. PROBLEM DESCRIPTION

From literature survey, it has been seen that there are various attempts being made towards addressing sentiment analysis. An early work on verifiable perspective location is [17], in that the creators propose to utilize semantic affiliation examination dependent on point-wise shared data (PMI) to separate certain perspectives from single notional words. Tragically, there were no quantitative trial results announced in their work, yet naturally the utilization of measurable semantic affiliation investigation ought to take into account certain assessment words, for example, "huge," to gauge the related angle ("size"). In [18], a methodology is recommended that at the same time and iteratively bunches item angles and assessment words. Perspectives/assessment words with high comparability are bunched together, and viewpoints/assessment words from various groups are different. In [19], a semi-solo technique is suggested that can at the same time separate both opinion words and item/administration viewpoints from audit sentences.

A. Proposed Solution

In this paper, both a solo and a regulated methods are recommended that can discover viewpoint classes dependent on co-event frequencies. The solo strategy utilizes spreading actuation on a diagram worked from word co-event frequencies to recognize perspective classifications.

The proposed unaided strategy utilizes something other than the strict class mark by making a bunch of unequivocal lexical portrayals for every classification.

The just required data is the arrangement of viewpoint classes that is utilized inthe informational index. The administered technique then again utilizes the co- events between words, just as syntactic connection significantly increases, and the commented on viewpoint classifications to ascertain contingent probabilities from which location rules are mined.

III. METHODOLOGY AND IMPLEMENTATION

A. *Unsupervised Method*

The proposed unsupervised method (called the spreading activation method) uses co-occurrence association rule mining in a similar way as [15], by learning relevant rules between notional words, defined as the words in the sentence after removing stop words and low frequency words, and the considered categories. This enables the algorithm to imply a category based on the words in a sentence. To avoid having to use the ground truth annotations for this and to keep this method unsupervised, we introduce for each category a set of seed words, consisting of words or terms that describe that category.

These words or terms are found by taking the lexicalization of the category, and its synonyms from a semantic lexicon like WordNet. For example, the ambience category has the seed set {ambience, ambiance, atmosphere}. With the seed words known, the general idea of implicit aspect detection can be exploited to detect categories as well. The idea is to mine association rules of the form [notional word → category] from a co- occurrence matrix. Each entry in this co-occurrence matrix represents the frequency degree of two notional words co-occurring in the same sentence. Stop words, like the and and, as well as less frequent words are omitted because they add little value for determining the categories in review sentences.

The reason why we choose to mine for rules similar to that of [15]'s, and do not consider all notional words in the sentence at once to determine the implied categories, like [21], is based on the hypothesis that categories are better captured by single words. If we have for example categories like food and service all it takes to categorize sentences is to find single words like chicken, staff, or helpful.

Association rules are mined when a strong relation between a notional word and one of the aspect categories exists, with the strength of the relation being modeled using the co-occurrence frequency between category and notional word.

We distinguish between two different relation types: 1) direct and 2) indirect relations. A direct relation between two words A and B is modeled as the positive conditional probability P(B|A) that word B is present in a sentence given the fact that word A is present.

B. *Supervised Method*

Similar to the first method, the supervised method (called the probabilistic activation method) employs co-occurrence association rule mining to detect categories. We borrow the idea from to count co-occurrence frequencies between lemmas and the annotated categories of a sentence. However, low frequency words are not taken into account in order to prevent overfitting. This is achieved using a parameter αL, similar to the unsupervised method. Furthermore, stop words are also removed.

As well as checking the co-events of lemmas and perspective classes, the co- events between syntactic conditions and angle classifications are likewise tallied. Like lemmas, low recurrence conditions are not considered to forestall overfitting, utilizing the boundary αD. Conditions, portraying the linguistic relations between words in a sentence, are more explicit than lemmas, as every reliance has three parts: 1) lead representative word; 2) subordinate word; and 3) connection type. The additional data given by conditions, may give more exact expectations, with regards to class location. Knowing whether a lemma is utilized in a subject connection or as a modifier can have the effect among anticipating and not foreseeing a class.

Once the conditional probabilities are computed and the thresholds are known, unseen sentences from the test set are processed. For each unseen sentence we check whether any of the lemmas or dependency forms in that sentence have a conditional probability greater than its corresponding threshold, in which case the corresponding category is assigned to that sentence. Fig. 3 illustrates how the supervised method works on a very simple test and training set.

IV. RESULTS AND DISCUSSION

For the evaluation of the proposed methods, the training and test data from SemEval-2014 [10] are used. It contains 3000 training sentences and 800 test sentences taken from restaurant reviews. Each sentence has one or more annotated aspect categories. Fig. 4.1 shows that each sentence has at least one category and that approximately 20% of the sentences have multiple categories. With 20% of the sentences having multiple categories, a method would benefit from being able to predict multiple categories. This is one of the reasons why association rule mining is useful in this scenario as multiple rules can apply to a single sentence.

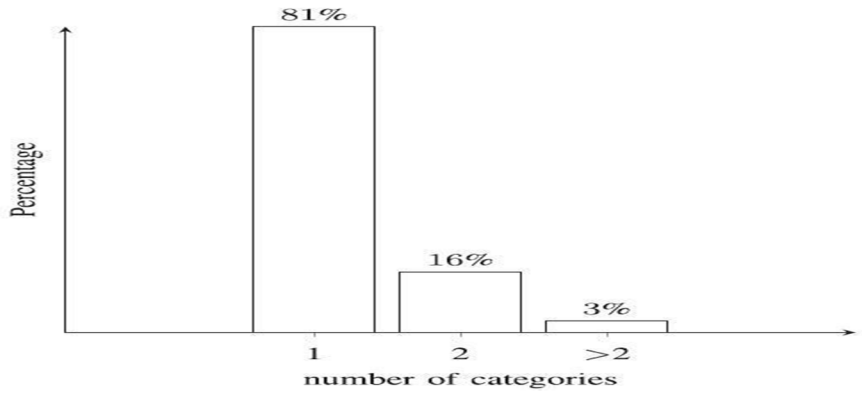


Fig. 4.1: Distribution of number of aspect categories per sentence.

Fig. 4.2 presents the relative frequency of each aspect category, showing that the two largest categories, food and anecdotes/miscellaneous, are found in more than 60% of the sentences. This should make these categories easier to predict than the other categories, not only because of the increased chance these categories appear, but also because there is more information about them.

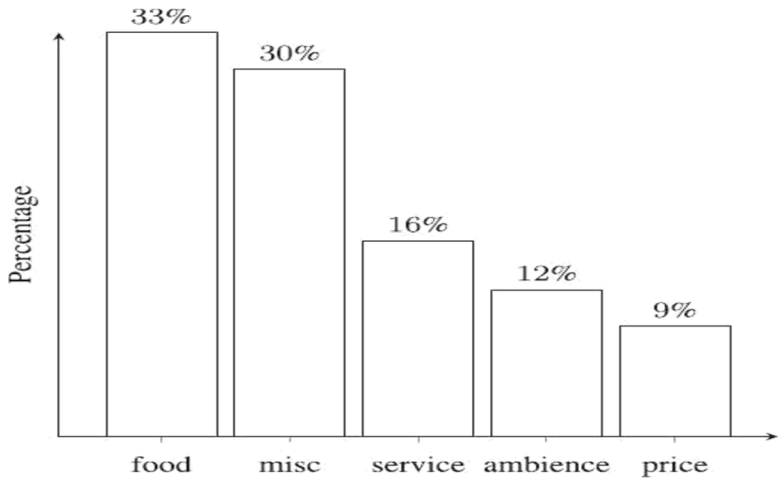


Fig. 4.2: Relative frequency of the aspect categories.

Last, in Fig. 4.3, the proportion of implicit and explicit aspect categories is shown. It is clear that using techniques related to implicit aspect detection is appropriate here, given that more than three quarters of the aspect categories is not literally mentioned in the text.

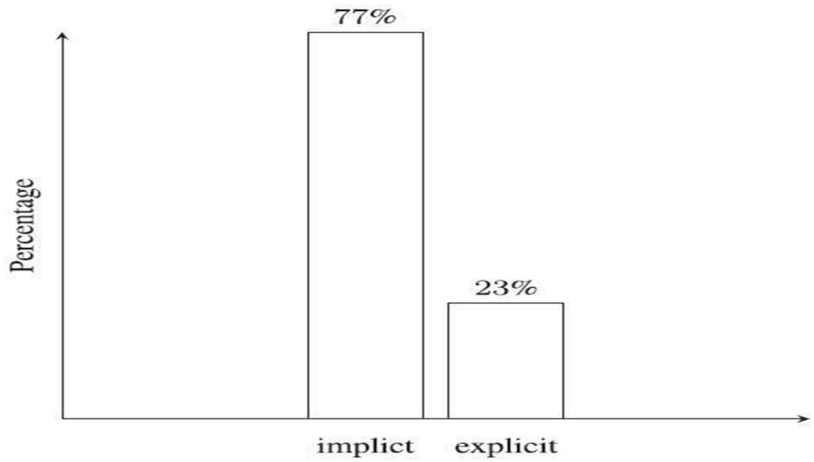


Fig. 4.3: Ratio between implicit aspect categories and explicitly mentioned ones.

Because both unsupervised and supervised method work best for well-defined aspect categories, the last category in this data set, anecdotes/miscellaneous poses a challenge. It is unclear what exactly belongs in this category, and its concept is rather abstract. For that reason, we have chosen not to assign this category using any of the actual algorithms, but instead, this category is assigned when no other category is assigned by the algorithm. The characteristics in Fig. 4.3 also show that the use of anecdotes/miscellaneous as a "fallback" is justified given its large size and the fact that every sentence has at least one category.

A. Unsupervised Method

Table I displays, for each aspect category, the chosen firing threshold together with the resulting precision, recall, and F1-score on the test set. The category anecdotes/miscellaneous is estimated when none of the other four categories are chosen in the sentence.
From Table I, one can conclude that this approach has difficulty predicting the category ambience. This might be due to the nature of that particular category, as it is often not specified in a sentence by just one word, but is usually derived from a sentence by looking at the sentence as a whole.

Table I. Chosen Firing Thresholds and Their Evaluation Scores on the Test Set

Category	TP's	FP's	FN's	τ_c	precision	recall	F_1
food	313	103	105	0.22	75.1%	74.4%	74.8%
service	100	4	72	0.19	96.2%	58.1%	72.5%
ambience	41	10	77	0.09	80.4%	34.8%	48.5%
price	52	16	31	0.09	79.0%	54.2%	64.3%
misc.	163	159	71	-	50.6%	70.9%	59.1%
all	852	157	173	-	70.0%	64.7%	67.0%

B. Supervised Method

For the supervised method we use the training set to learn the parameters and co- occurrence frequencies, after which we evaluate the method on the test set. To see the impact the dependency indicators have, this method is executed separately for the dependency indicators, lemma indicators and a combined version where both lemma and
dependency indicators are used, and evaluated on the test set. Tables II–III show the results.

Table II. Evaluation Scores of the Supervised Method with both Dependency and Lemma Indicators on the Test Set

Category	TP's	FP's	FN's	precision	recall	F_1
food	371	51	47	87.9%	88.8%	88.3%
service	159	32	13	83.2%	92.4%	87.6%
ambience	83	28	35	73.8%	70.3%	72.5%
price	74	8	9	90.2%	89.2%	89.7%
anecdotes/misc.	165	38	69	81.3%	70.5%	75.5%
all	852	157	173	84.4%	83.1%	83.8%

Table III. Evaluation Scores of the Supervised Method with Only Dependency Indicators on the Test Set

Category	TP's	FP's	FN's	precision	recall	F_1
food	343	45	75	88.4%	82.1%	85.1%
service	152	27	20	84.9%	88.4%	86.6%
ambience	62	34	56	64.6%	52.5%	57.9%
price	61	5	22	92.4%	73.5%	81.9%
anecdotes/misc.	165	38	69	81.3%	70.5%	75.5%
all	783	149	242	84.0%	76.4%	80.0%

The two techniques introduced for recognizing angle classes that is helpful for online audit rundown. The first, unaided, strategy, utilizes spreading enactment over a chart worked from word co-event information, empowering the utilization of both immediate and aberrant relations between words. This outcomes in each word having an actuation an incentive for every classification that addresses that it is so liable to suggest that classification.

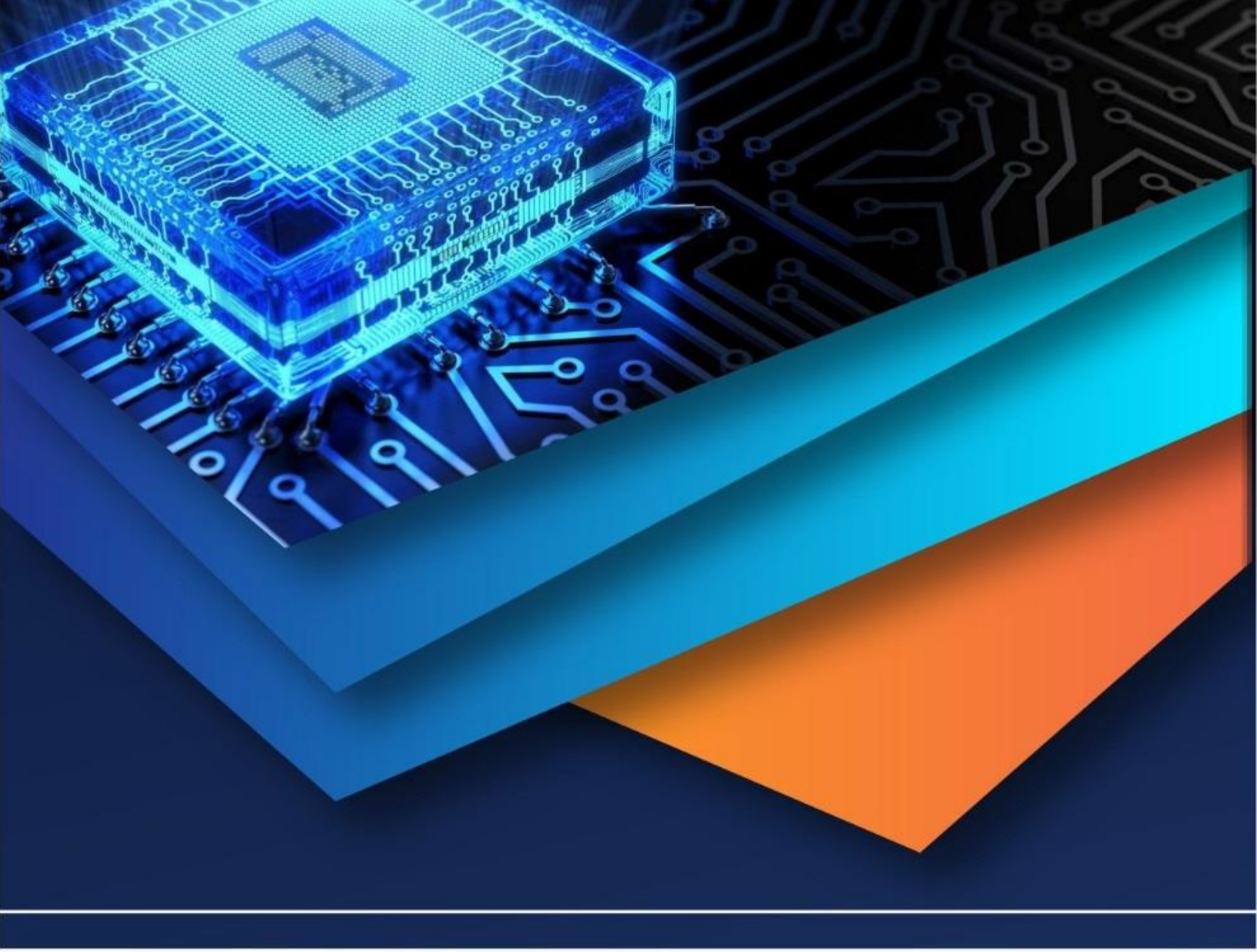
While different methodologies need marked preparing information to work, this strategy works solo. The significant downside of this strategy is that a couple of boundaries should be set already, and particularly the classification terminating edges (i.e., τc) should be painstakingly set to acquire a decent exhibition. We have given heuristics on how these boundaries can be set.

The second, administered, technique utilizes a fairly direct co-event strategy where the co-event recurrence between commented on viewpoint classes and the two lemmas and conditions is utilized to ascertain restrictive probabilities. On the off chance that the greatest contingent likelihood is higher than the related, prepared, edge, the classification is alloted to that sentence. Evaluating this approach on the official SemEval-2014 test set [10], shows a high F1-score of 83%.

REFERENCES

[1] P. F. Bone, "Word-of-mouth effects on short-term and long-term product judgments," J. Bus. Res., vol. 32, no. 3, pp. 213–223, 1995.
[2] R. Feldman, "Techniques and applications for sentiment analysis," Commun. ACM, vol. 56, no. 4, pp. 82–89, 2013.
[3] S. Sen and D. Lerman, "Why are you telling me this? An examination into negative consumer reviews on the Web," J. Interact. Marketing, vol. 21, no. 4, pp. 76–94, 2007.
[4] B. Bickart and R. M. Shindler, "Internet forums as influential sources of consumer information," J. Consum. Res., vol. 15, no. 3, pp. 31–40, 2001.
[5] D. Smith, S. Menon, and K. Sivakumar, "Online peer and editorial recommendations, trust, and choice in virtual markets," J. Interact. Marketing, vol. 19, no. 3, pp. 15–37, 2005.
[6] M. Trusov, R. E. Bucklin, and K. Pauwels, "Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site," J. Marketing, vol. 73, no. 5, pp. 90–102, 2009.
[7] M. T. Adjei, S. M. Noble, and C. H. Noble, "The influence of C2C communications in online brand communities on customer purchase behavior," J. Acad. Marketing Sci., vol. 38, no. 5, pp. 634–653, 2010.
[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.
[9] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou, "Movie rating and review summarization in mobile environment," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 42, no. 3, pp. 397–407, May 2012.
[10] 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.
[11] S. Kiritchenko, X. Zhu, C. Cherry, and S. M. Mohammad, "NRCCananda- 2014: Detecting aspects and sentiment in customer reviews," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 437–442.
[12] 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.
[13] C. R. C. Brun, D. N. Popa, and C. Roux, "XRCE: Hybrid classification for aspect- based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 838–842.
[14] G. Castellucci, S. Filice, D. Croce, and R. Basili, "UNITOR: Aspect based sentiment analysis with structured learning," in Proc. 8 Int. Workshop Semantic Eval.
[15] (SemEval), Dublin, Ireland, 2014, pp. 761–767.
[16] 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.
[17] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," IEEE Trans. Knowl. Data Eng., vol. 28, no. 3, pp. 813–830, Mar. 2016.
[18] 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.
[19] 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.
[20] 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,"
[21] 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.
[22] 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.
[23] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," Comput. Linguist., vol. 37, no. 1, pp. 9–27, 2011.
[24] K. Schouten, F. Frasincar, and F. de Jong, "COMMIT-P1WP3: A co-occurrence based approach to aspect-level sentiment analysis," in Proc. 8th Int. Workshop
[25] 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.
[26] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proc. 32nd Annu. Meeting Assoc. Comput. Linquistics, Las Cruces, NM, USA, 1994, pp. 133–138.
[27] [26]F. Crestani, "Application of spreading activation techniques in information retrieval," Artif. Intell. Rev., vol. 11, no. 6, pp. 453–482, 1997.
[28] S. Bagchi, G. Biswas, and K. Kawamura, "Task planning under uncertainty using a spreading activation network," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 30, no. 6, pp. 639–650, Nov. 2000.
[29] A. Katifori, C. Vassilakis, and A. Dix, "Ontologies and the brain: Using spreading activation through ontologies to support personal interaction," Cognitive Syst. Res., vol. 11, no. 1, pp. 25–41, 2010.

[30] 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
[31] M.-C. de Marneffe and C. D. Manning, "Stanford typed dependencies manual," Stanford NLP Group, Stanford University, Stanford, CA, USA, Tech. Rep., Sep. 2008. [Online]. Available: https://nlp.stanford.edu/software/dependencies_manual.pdf
[32] 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY