EasyChair Preprint

№ 5162

# Predicting Supervised Machine Learning Performances for Sentiment Analysis Using Contextual Based Approaches

Venkata S Lakshmi, K Janan, Joshua P S Joseph and Mohammed Sharoz



March 16, 2021

# PREDICTING SUPERVISED MACHINE LEARNING PERFORMANCES FOR SENTIMENT ANALYSIS USING CONTEXTUAL BASED APPROACHES

**DR. S VENKATA LAKSHMI [1,2], JANAN K [1,3], JOSHUA JOSEPH P S [1,4] AND MOHAMMED SHAROZ [1,5]**

[1]Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India.

[2]venkatalakshmis@skcet.ac.in, [3]17eucs065@skcet.ac.in, [4]17eucs069@skcet.ac.in, [5]17eucs096@skcet.ac.in

***ABSTRACT*** The fundamental thought of our methodology is to inspire client inclinations communicated in text-basedaudits, an issue known as opinion investigation, and guide such inclinations onto some evaluating scales that can beperceived by existing CF calculations. One significant errand in our rating deduction system is the assurance of wistfuldirections PSWAM and qualities of assessment words. It is because surmising a rating from an audit is chiefly done by separating assessment words in the survey, and afterward accumulating the PSWAM of such words to decide the predominant or normal notion inferred by the client. We played out some primer examination on film audits to research how PSWAM and qualities of assessment words can be resolved and proposed a relative- a recurrence-based technique for performing such assignments. The proposed technique tends to a significantimpediment of existing strategies by permitting comparative words to have distinctive PSWAM. We additionally created and assessed a model of the proposed structure. Fundamental outcomes approved the viability of different assignments in the proposed system and suggested that the process doesn't rely on a large preparation corpus for working. An accelerated algorithm based on the Naïve Bayes approach is used to solve the PSWAM and a parallel algorithm based on FISTA is incorporated to further improve the efficiency. The result is a graph representing opiniontarget and opinion word candidates before and after extraction further helping users simplify the task of analysis.

***KEYWORDS Partially-Supervised Word Alignment Model, Fast Iterative Shrinkage-Thresholding Algorithm, Opinion Mining, Naive Bayes***

## 1. INTRODUCTION

Mining the supposition data in the monstrous client-created substance can assist with detecting thepopular sentiments towards a combination of points,like items, brands, catastrophes, occasions,superstars thus going on, and it is helpful in numerous applications. For events, specialists haveset up that examining the notions in tweets has thelikely to predict qualification of stock commercial center costs and official choice outcomes. Grouping the estimations of gigantic miniature blog messagesare likewise useful to substitute or enhancecustomary surveying, which is costly and tedious. Item survey opinion examination can assist organizations with improving their items and administrations, and help clients settle on more educated choices. Investigating the assessments of client produced fulfilled is additionally affirmed valuable for customer premium evacuation, customized proposal, social exposure, buyer connection theexecutives, and emergency the board. Thus, notiongrouping is a hot exploration point in both modern and scholarly fields.

In some larger part supposition study strategies, theestimation plan is viewed as a section grouping issue. Regulated AI methods, like SVM, Logistic Regression, and CNN, are regularly applied to prepare feeling classifiers on named datasets and anticipate the assumptions of concealed writings. These techniques have been utilized to investigate the opinions of item audits, miniature

websites. Then again, assumption arrangementis generally perceived as an area subordinate issue.This be divergent spaces present are diverse reaction words, and the equivalent word could propose bizarre notions in various areas. For instance, in the area of electronic item surveys"simple" is typically certain. However, in the space of film audits, "simple" is regularly utilized as anegative word. Thus, the supposition classifier prepared in one space may neglect to catch the particular slant articulations of another area, and its presentation in an alternate area is normally unsuitable.

An unstructured answer for this difficulty is to directan area point by point supposition classifier for both spaces with the named tests of this field. In any case, the named information in numerous spaces isoften scant. As present are tremendous areas involved in online client-created content, it is expensive and long to clarify enough examples for them. Without sufficient marked information, it is genuinely hard to show a right and good space explicit assessment classifier for every region self- sufficiently. The inspiration of our work is that albeitevery space has its particular conclusion articulations, various areas likewise share numerous normal assessment words.

This work prepares conclusion classifiers for different areas all the while in a cooperative manner.In this methodology, the notion classifier of every area is disintegrated into two segments, i.e., a worldwide one and space explicit one. The area explicit inclination classifier is shown utilizingmarked examples of one space and can detain the space explicit disposition articulations. The worldwide supposition classifier is shared by allspaces and is prepared on the marked examples from different areas to have better speculation capacity. It can catch the overall opinion informationin predictable unconcerned areas. Also, separate earlier broad conclusion information fromuniversally useful opinion dictionaries and consolidate it into our way to deal with control the learning of the worldwide notion classifier.Additionally, propose to remove area explicit slant information for every space from both restrictednamed tests and monstrous unlabelled examples. The area explicit conclusion information is utilized toupgrade the learning of space explicit assessment classifiers in approach. Two sorts of area similitudemeasures are investigated, one dependent on the text-based substance, and the other one dependent on thefeeling word conveyance.

## 2. RELATED WORKS

A critical piece of our investigation conduct has forever been to discover extra's opinion. With the developing accessibility and notoriety of assessment rich capital, for example, online surveylocales and private websites, new freedoms, and difficulties happen as individuals currently can, anddo, forcefully use in grouping innovations to look for out and perceive the assessments of others. The unforeseen blast of action nearby view miningand slant study, which manages the computational treatment of assessment, estimation, and partisanship in text, has subsequently happened atany rate in component as an explicit reply answer to the surge of revenue in imaginative frameworks that manage suppositions as an unmatched item. This has been explained by B. Ache and L. Lee [1]in their paper. Another approach as explained by B. Liu [2], is by utilizing a psychometric machine to eliminate the six demeanor states (pressure, discouragement, bothering, imperativeness, weariness, vulnerability)from the collected dataset and work out a six-dimensional temper vector for every day in the course of events.

The work was further upgraded by J Bollen, H. Mao [3], interfacing activities of generalassessment exact from the surveys with feelings determined from the text. While the result changes across datasets, in a significant number of cases the relationships are pretty much as high as 80% and catch basic huge scope patterns. Chen, R. Xu, Y [6] in their research, have arrangedis basic for some dissimilar to applications, for example, the executives and industry insight to explore and stroll around the spread of popularfeelings via website-based media. However, the quick multiplication and incredible arrangement of public assessments via online media presentextraordinary difficulties to effective assessment of the scattering process.

To start a visual report framework called assessment pour to permit examiners to see assessment dissemination designs and gather bits of knowledge. Mixed with thein-arrangement circulation model and thepresumption of specific disclosure, build up an assessment dispersal multiplication to ballpark assessment broadcast among Twitter clients. The work by Y. Wu, S. Liu, K. Yan [7], has proposed to consider the issue of group reports not by subject, but rather by for the most part assessment, e.g., persuasive whether an audit is idealistic or apathetic.

Utilizing film audits as information, noticethat standard motor learning methods completely show improvement over human-created baselines. All things considered, the three AI techniques are locked in (Naive Bayes, most noteworthy entropy classifying, and uphold vector machines) don't make too on estimation grouping as on since quite a while ago settled point-based marking system. To end byconditional components that makes the feeling arrangement issue seriously requesting. The work that was proposed by E. Cambria [9], on microblog disposition order is a focal explore point which has wide applications in both the scholarly world and industry. Since miniature blog messagesare short, uproarious, and contain masses of abbreviations and casual words, miniature blog opinion order is exceptionally difficult to undertake. Propitiously, together the relevant data about these peculiar words give information about their estimation directions. The research proposes to utilize the miniature web journals' relevant information mined from a lot of unlabelled information to help improve miniature blog notion grouping.

It depicts two sorts of foundation information: explanation association and word-estimation association. The work done by B. Ache, L. Lee [10], has arranged mechanical feeling association that has been lengthily determined and utilitarian in a new time. Then again, the supposition is articulated in analternate path in different areas, and clarifying corpora for each likely zone of consideration isn't suitable. To investigate area release for feeling classifiers, zeroing in on online audits for unique kinds of products. In the first place, stretch to reaction game plan the as of late proposed underlying correspondence learning (SCL) calculation, dropping the connection shortcoming because of version between areas by a normal of 30% over the first SCL calculation and 46% over a managed pattern. Second, to recognize aproportion of space examination that relates well with the feasibility for the transformation of a classifier starting with one territory then onto the next. This ascertain could for example be utilized tochoose a little arrangement of areas to explain whose trained classifiers would move well to numerous different spaces.

## 3. PROPOSED METHODOLOGY

Two sorts of information are consolidated to removearea explicit supposition information for every space. The primary sort of information is the namedtests, which are related to opinion marks and can beutilized to construe area explicit conclusion articulations straightforward manner. A typical perception in the notion examination field is that the words happening more often in sure examples than negative examples normally will in a general pass on sure conclusion directions, and the other way around.

In this way, we can proliferate the feeling marks from archives/sentences to words to separate the area's explicit assessment articulations. A few pre-processing steps were taken before tests. Words were changed over to bring down cases and stop words were taken out. In this paper, we propose to remove the underlying estimation scores of words depending on their dissemination contrasts in certain and negative examples.

### 3.1 PSWAM MODEL

Given various areas to be examined, few named tests in these spaces, the area likenesses between them, the overall feeling information separated from broadly useful assumption dictionaries, and the area explicit

slant information on every area extricated from both marked and unlabelled examples, the objective of our methodology is to prepare precise space explicit slant classifiers for numerous spaces in a cooperative manner.

### 3.2 FISTA ALGORITHM

FISTA based sped-up calculation for our methodology which can be led on a solitary figuring hub. As referenced previously, the advancement issue in our methodology is not smooth. Despite the fact that we can utilize sub slope plunge technique to tackle it, the union pace of the sub inclination strategy is O(1/√k) and is a long way from palatable, where k is the quantity of emphasis. In this manner, we propose to utilize the sped-up calculation dependent on FISTA to take care of the streamlining issue. When f is smooth, (for example, squared misfortune and log misfortune). This calculation has a similar computational intricacy as slope strategy and subgradient technique in every cycle and simultaneously has a combination pace of O(1/k^2) a lot quicker than that of inclination technique (O(1/k)) and sub angle technique (O(1/√k)).

.

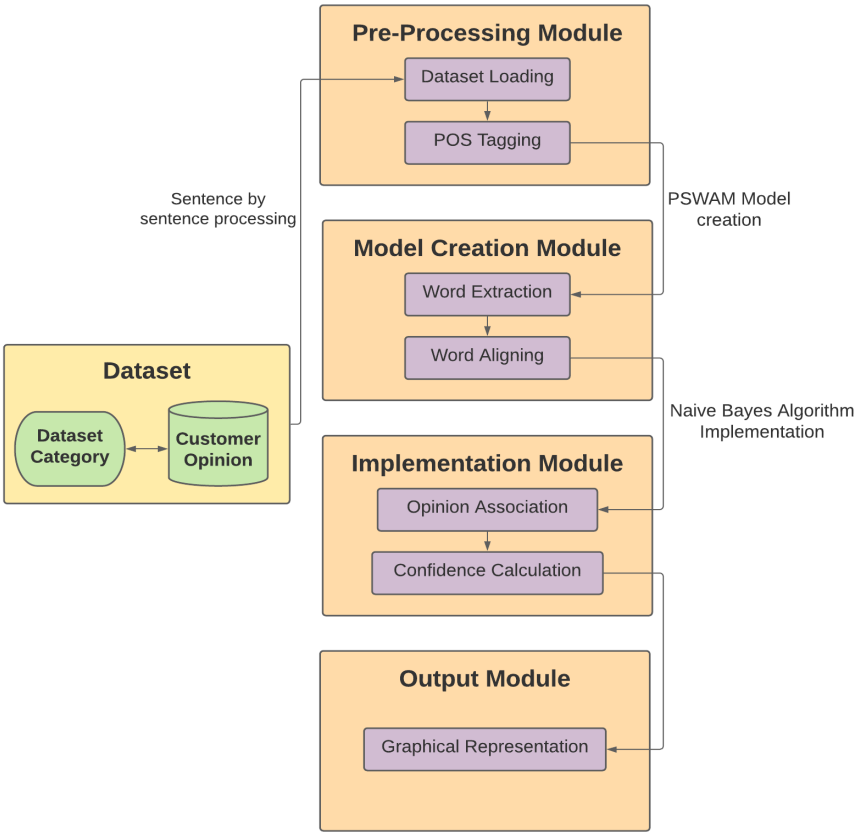


**Figure - 1 System Architecture**

## 4. MODULES

### 4.1 Preparing the dataset for processing

The provided dataset is loaded onto the program. The dataset provided for this example is based on the customer's review of an electronic product. The dataset that is loaded is now shown to the user for confirmation of the right dataset. After loading the next step is to process the data that was uploaded. All these processes are accomplished by using the library called "stanford_postagger" that processes the dataset where each part of speech that present in the dataset is tagged. The data undergoes Natural Language Processing to separate the nouns, verbs, numbers, and other parts-of-speech.

.

**Table - 1 POS TABLE**

| Tags | Parts-Of-Speech |
|---|---|
| /NN | NOUN |
| /VB | VERB |
| /NNB | PRONOUN |
| /JJ | ADJECTIVE |
| /CD | DECIMAL |
| /IN | PREPOSITION |

Once the dataset is loaded with the help of the browse function used for directory traversal, the dataset then undergoes the process of opinion mining where the opinion from the user is separated into sentences for further Parts of speechTagging.

**4.2 Extracting candidate words and preparing a PSWAM model:**

A Partial-Supervised Word Alignment Model is created. PSWAM is most often used in sentences and is used for estimating the relation betweenwords for mining opinion relations. The dataset is divided into sentences that are further divided by their separation using commas. The different types of nouns such as plural nouns, possessive nouns are broadly classified as nouns and are categorized as Opinion Target Candidates. Similarly, the different types of verbs and adjectives are categorized as Opinion Word Candidates. The Processed data is then displayed in a table to the user. A table containing the opinion targets and opinion words that are present in a sentence is created. The next step of the process is by aligningthe words we extracted by separating each of the nouns and verbs into separate Opinion Targets and Opinion Word candidates. Each opinion target word is associatedwith a corresponding opinion word and it is displayed in the form of a table to the user.

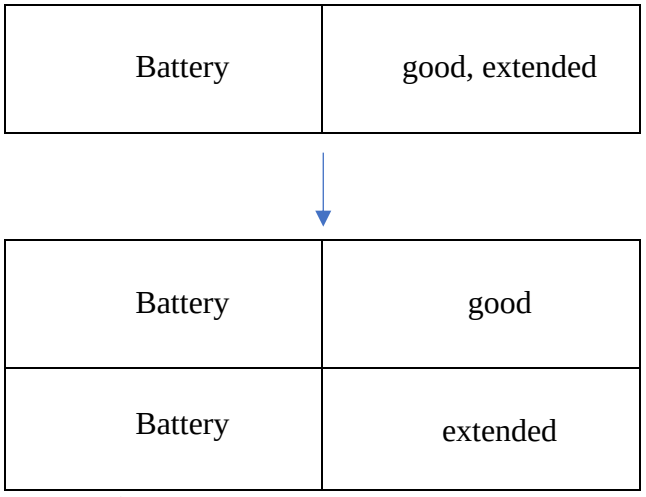


**Figure - 2 PSWAM MODEL CREATION**

The Naive Bayes method is actually a technique used for classification based on Bayes' theorem by assuming there is independence among the predictors. To be clear, the classifier that works based on Naïve Bayes considers that one particularfeature present in a class is not in any way related to any other feature that is present in a class. One advantage of Naive Bayes is that it can be built so easily and also can be used for huge data sets. Naive Bayes is not only used for its simplicity but themain interesting feature is that its performance is farbetter when compared to many other complex methods of classification.

.

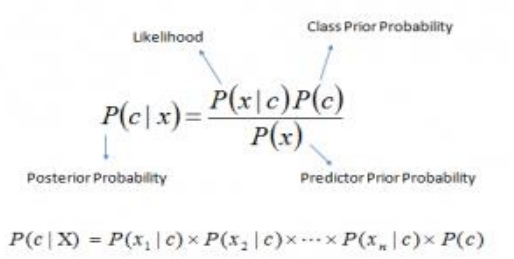


**Figure - 3 NAIVE BAYES EQUATION**

Most often, in the process of reviewing a sentence,the product Features are nouns or noun phrases. The collected data or user-reviews are sentences ortext. The Natural Language Processor makes use of the linguistic parser which is used to divide the data into sentences according to constraints such as punctuation and in turn creates parts of speech tag suitable for the sentences based on word type such as noun, verb, adjective. The result is nouns and vers are grouped. The word alignment model uses two constraints one is the opinion target candidates which includes nouns andnoun phrases and the rest of the adjectives and verbcomes under opinion word candidates. Finally, the opinion target words and candidate words are aligned separately by the partially supervised model.

**4.3 Calculating the Opinion Association:**

The opinion association between the opinion target candidates and the opinion word candidates is calculated by formulating the alignment probability between an opinion targe($w_t$) and the opinion word($w_o$). It is estimated using,

$$P(w_t|w_o) = \frac{Count(w_t, w_o)}{Count(w_o)} \quad - \quad (1)$$

The alignment probability between an opinion word($w_o$) and the opinion target($w_t$) is estimated using,

$$P(w_o|w_t) = \frac{Count(w_t, w_o)}{Count(w_t)} \quad - \quad (2)$$

The opinion association value between the target candidates and word candidates is calculated using the alignment probabilities of opinion targets and opinion word candidates. It is estimated using the formula,

$$OpinionAssociation(w_t|w_o) = \left(\alpha * P(w_t|w_o)\right) + (1-\alpha)P(w_o|w_t)^{-1} \quad - \quad (3)$$

Here the α is the harmonic factor between two words. We take the value of α as 0.5.

**4.4 Computing confidence and finding opinion target and opinion words:**

The confidence of each opinion target and opinion words are calculated using the Random Walk method. The initial confidence for the opinion targetand opinion words are assumed as a value between0 and 1. The Random Walk Method uses the formula,

$$Confidence_t^{k+1} = (1-\mu) * OA_{to} * Confidence_o^k + \mu * I_t$$

$$Confidence_o^{k+1} = (1-\mu) * OA_{to} * Confidence_t^k + \mu * I_o \quad - \quad (4)$$

Here μ takes either value 0 or 1. If μ = 1 then theconfidence of the candidate is determined by prior knowledge. If μ = 0 then the confidence is determined candidate opinion relevance. $It$ and $Io$ isa score that denotes prior knowledge of the candidates being opinion targets and opinion words. We use a library "Sentiwordnet" to determine the score for the prior knowledge. $OA_{to}$ is the opinion association score that we calculate in the prior module. k represents the iteration count for thetargets and words. The calculated values are displayed to the user in the form of a table. We next calculate the target threshold and world thresholdfor the confidence values

We will sort the list containing the confidence valuesand choose the value at the middle to be the threshold value. The values greater than the threshold values are then chosen as the opinion targets and opinion words. The list is then displayed to the user in the form of a table. The confidence of the target candidate and word candidates are shown in Figure(4).

Target Threshold: 3

Word Threshold: 0.4001

| Opinion Target | Opinion Target Confide... | Opinion Words | Opinion Word Confiden... |
|---|---|---|---|
| life | 83.0688 | standard | 22.9481 |
| battery | 7.3077 | long | 2.0188 |
| battery | 560 | awesome | 0.75 |
| life | 4.6839 | long | 1.294 |
| life | 27.7541 | awesome | 7.6672 |
| battery | 560 | learned | 0.4091 |
| life | 258 | learned | 0.4091 |

**Figure - 4 OPNION TARGEST AND OPNION WORDS**

### 4.5 Preparing a graphical representation:

The number of opinion targets and opinion words before and after the extraction are displayed to the user in the form of a bar graph. The bar graph is plotted vertically by using two values. Mainly the number of opinion targets candidates and opinion word candidates and the number of opinion targetsand opinion words that are co-extracted.The colorfor the bars in the graph is set and then the graph isdisplayed to the user. The bar graph is created using the "jfreechart" library. The difference between the initial candidates and the final targets is given in Figure – 5.

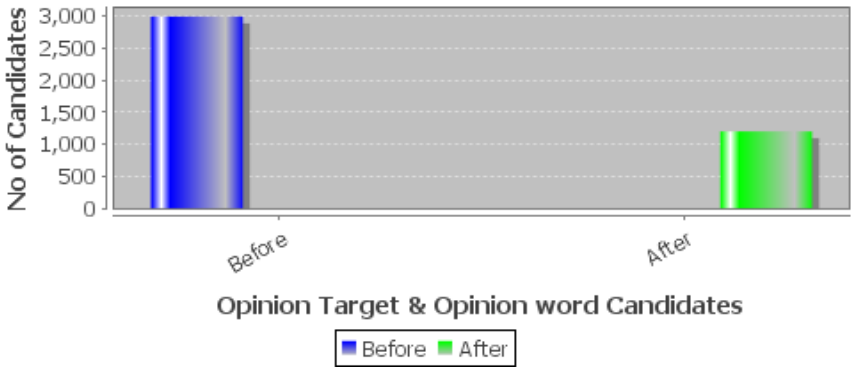


**Figure - 5 The graph representing opinion target and word candidates before and after extraction**

## 5. EXPERIMENTAL SETUP

### 5.1 Datasets and experimental settings

Two benchmark multi-area slant datasets were utilized in our tests. The first is the renowned Amazon item audit notion dataset1 (signified as which was gathered by Blitzer et al.also, incorporates four spaces, i.e., Book, DVD, Electronics, etc. It is generally utilized in multi-space and cross-area supposition examination fields. In every area, 1,000 positive and1,000 negative audits are incorporated. The second dataset was likewise slithered by Blitzer et al. from Amazon.

### 5.2 Comparison of domain similarity measures

In this part, we led investigations to sort out which one of the two area closeness measures presentedis more reasonable for the multi-space assessmentcharacterization task. The exploratory outcomes onthe Amazon-4 dataset that appear in Fig. 5.3 and the outcomes on the Amazon-21 dataset show comparative examples. Pivot misfortune was utilized in our methodology in these tests. The presentation of our methodology with various types of space similitude. NoSim, ContentSim, and SentiSim address the presentation of our methodology with no area comparability, with printed content-based space similitude, and with assumption articulation-based area closeness in a separate manner. The distinction between SentiSim-Initialand SentiSim-Prop is that the previous depends onthe underlying assumption scores extricated from named tests, and the last depends on the assessment scores after engendering

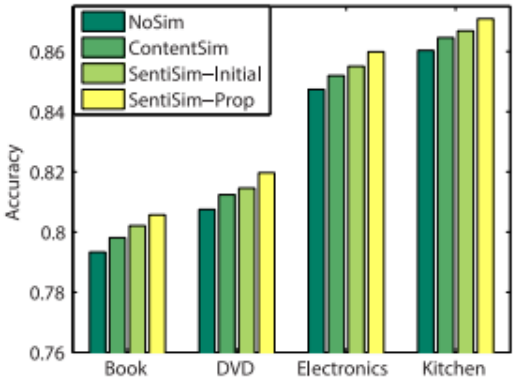


**Figure (5.3) KINDS OF DOMAIN SIMILARITY**

From Figure 5.3, we can see that the presentation of our shared multi-space conclusion groupingapproach with notion articulation-based area likeness is superior to that with literary substance-based space similitude. This outcome shows that the area likeness dependent on feeling articulationscan more readily gauge the opinion relatednessbetween unexpected spaces in comparison to that dependent on the literary substance in multi-area estimation characterization task.

### 5.3 Time Efficiency

We led a few trials to investigate the time intricacy of our methodology. The calculations were executed utilizing MatLab 2014a. All analyses weredirected on a workstation with Intel Core i7 CPU (3.4GHz) and 16 GB RAM. The single-hub adaptation FISTA-put together sped up calculation was led withrespect to a solitary center of this machine, and theADMM-based equal calculation was dispersed across the 4 centers of this machine. The trials wereled on the Amazon-21 dataset. In each analysis, wehaphazardly chose r of the named tests in every area for preparing. We changed the proportion r from 5 to 50 percent.

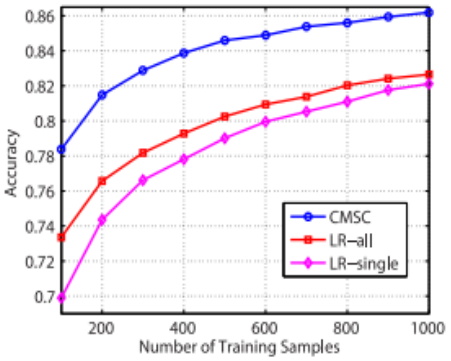


**Figure 5.3(a) DIFFERENT NUMBERS OF TRAINING SAMPLES**

Fig. 5.3(a). The normal presentation of our methodology and benchmark techniques on the fourareas of the Amazon-4 dataset with variousquantities of preparing tests. CMSC addresses our shared multi-space

feeling grouping approach. LR- all and LR-single are Logistic Regression opinion classifiers prepared on completely marked examples and single-space named tests,separately.

We can see that the running season of our methodology with various types of misfortune capacities is around straight regarding the size of the preparation information. This outcome approves our investigation of the time intricacy. Furthermore, our methodology with log misfortune (CMSC-Log) and squared misfortune

(CMSC-LS) runs a lot quicker than that with pivot misfortune (CMSC-SVM). It approves the convenience of the sped-up calculation dependent on FISTA in improving the productivity of our methodology. Furthermore, the running season of the equal calculation is fundamentally not exactly that of single-hub form enhancement calculation. It approves the adequacy of our equal calculation in accelerating the learning interaction via preparing conclusion classifiers for numerous spaces in equal at various processing hubs.

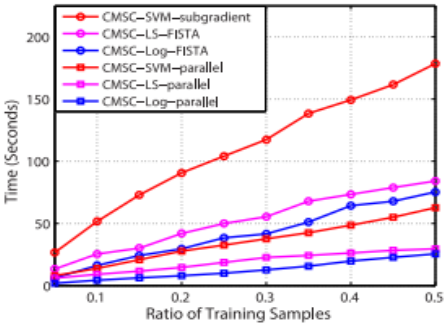


**Figure 5.3(b) DIFFERENT RATIO OF TRAINING SAMPLES**

## 6. CONCLUSION AND FUTURE PROSPECTS

This paper proposes a neoteric global model whichcan capture the general sentiment knowledge shared by different domains and the domain-specific models used to capture the specificsentiment expressions of each domain. Besides, weuse the prior general sentiment knowledge in general-purpose sentiment lexicons to guide thelearning of the global sentiment classifier .we also suggest implementing the similarities between various domains into our method by the process of sharing the sentiment data between similar domains with the help of regularization over the domain-specific sentiment classifier.

The approachis formulated into a convex optimization problem. Experimental benchmark results show that our method can effectively improve the performance of multi-domain sentiment classification, and significantly outperform baseline methods. In the future, we can integrate Big-data analytic tools to further improve the initial tagging and modelcreation in a much more efficient way compared to a manual creation.

## 7. REFERENCES

[1]B.Pang and L.Lee.Opinion Mining and Sentiment Analysis.Foundations and Trends in Information Retrieval January 2008 https://doi.org/10.1561/1500000011

[2]Bollen, J., Mao, H., & Pepe, A. (2011). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. Proceedings of the International AAAI Conference on Web and Social Media.https://ojs.aaai.org/index.php/ICWSM/article /view/14171

[3]B.O'Connor, R.Balasubramanyan, B.R. Routledge, and N.A.Smith, "From tweets to polls:
Linking text sentiment to public opinion time series," in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010

[4]Ye, Q., Zhang, Z., & Law, R. (2019). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, doi:https://doi.org/10.1016/j.eswa.2019.07.035

[5]Vinodhini, G., & Chandrasekaran, R. (2017). A sampling-based sentiment mining approach for e-commerce applications, Information Processing andManagement, Vol 53, 223-236, doi:https://doi.org/10.1016/j.ipm.2019.08.003

[6]Zhang, X., & Zheng, Z. (2019). Comparison of text sentiment analysisbased on machine learning. Paper presented at the 2019 18th International Symposium on Parallel and Distributed Computing (ISPDC), 230-233. doi:10.1109/ISPDC.2019.39
[7]S. Mahalakshmi and E. Sivasankar, ‘‘Cross-domain sentiment analysis using different machine learning techniques,’’.Fuzzy Neuro Comput. (FANCCO), 2015, pp. 77–87.
[8]A. A. Aziz, A. Starkey, and M. C. Bannerman, ‘‘Evaluating cross-domain sentiment analysis using supervised machine learning techniques,’’ in Proc. Intell. Syst. Conf. (IntelliSys), Sep. 2017, pp. 689–696, doi:10.1109/ intellisys.2017.8324369.
[9]B. Liu, Morgan & Claypool Publishers. Morgan & Claypool Publishers, 2012.
[10]M. M. Mirończuk and J. Protasiewicz, ‘‘A recent overview of the state-of-the-art elements of text classification,’’ Expert Syst. Appl., vol. 106, pp. 36– 54, Sep. 2018, doi:10.1016/j.eswa.2018.03.058.
[11]R. Feldman, ‘‘Techniques and applications for sentiment analysis,’’ Com- mun. ACM, vol. 56, no. 4, p. 82, Apr. 2013, doi:10.1145/2436256.2436274.
[12]T. Chen, R. Xu, Y. He, and X. Wang, ‘‘Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN,’’ Expert Syst. Appl., vol. 72,pp. 221-230, Apr 2017,doi:10.1016/j.eswa.2016.10.065
[13]A. Bagheri, M. Saraee, and F. de Jong, ‘‘An unsupervised aspect detection model for sentiment analysis of reviews,’’ in Proc. Natural Lang. Process. Inf. Syst., 2013, pp. 140–151.
[14]M. Tubishat, N. Idris, and M. A. Abushariah, ‘‘Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges,’’ Inf. Process. Manage., vol. 54, no. 4, pp 545-563, Jul. 2018, doi:10.1016/j.ipm.2018.03.008