International Journal of Engineering & Technology, 7 (4.5) (2018) 257-261

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



A Machine Learning Approach to Extract Opinions from Social Media Content

Salina Adinarayana[1]*, E Ilavarasan[2]

[1]Shri Vishnu Engineering College for Women, Bhimavaram, India
[2] Pondicherry Engineering College, Pondicherry, India
**Corresponding author E-mail:s.suhasini2k9@gmail.com*

Abstract

The Opinion Mining (OM) from mobile based social media content (SMC) is more challenging compared to topic-based mining, and it cannot be performed based on just examining the presence of single words in the text containing opinion expressions. Moreover, the existing systems of opinion classification find that a large number of features that are not feasible for the mobile environment. The existing methods of OM in this mobile environment do not consider the semantic orientation of the SMC in the review. The proposed machine learning approach extends the feature-based classification approach to identify the orientation of the phrase on taking context into account to improve the accuracy.

Keywords: *LDA, Latent Dirichlet Allocation, Opinion Mining (OM), CISDL, Mobile platform, social media content*

1. Introduction

Now-a-days, people are becoming more dependent on mobile gadgets in daily life, and there are numerous applications. In a mobile platform, it is not possible to show the complete review because of its screen size limitation. Therefore, there is a requirement for mining these reviews into two major categories such as positive and negative for ensuring the users an overview to make quick decisions on that particular review or subject. OM is a Natural Language Processing (NLP) that estimates the attitude, feelings, and opinions with respect to a particular topic, products, and services.

In mobile apps while classifying product review, movie review the whole review can be considered as document and at level, each document express some view, opinions. Documents don't represent a [16] single point of view, a single opinion. In this level it has multiple opinions representing several closely related opinions, as a result of document classification opinions will be classified into positive or negative class. In movie or blog review or product review every sentence consists of opinions this can be considered as Sentence stage classification, in which it classifies opinions as positive, negative or neutral class. Last stage in this sentiment classification is Entity stage where in which opinions features are identified from the source data. The importance of scientific and business communities increases in garnering public opinion in particular domains like political movements, social events, product preferences, company strategies, and marketing campaigns. There are several works have been surveyed on sentiment classification and OM [1][2].

In recent days, the area of OM has gained more interest [3] [4] [5]. Opinion mining is an automated collection of subjective content from the text and detects the orientation of the text as positive, negative, and neutral. The primary objective of opinion mining is polarity detection. There is a need to classify the polarity into two opposing sentiments if a segment of the text expressing an opinion on a single issue. Opinions such as "like" or "dislike" are instances of polarity classification. Polarity classification detects positive and negative expressions in online reviews and assists to evaluate the credibility of the product. Opinion summarization is necessary due to the limited size of the digital display of the mobile phones [6]. The opinion mining in a mobile environment involves two major steps. The first step classifies whether a sentence in a conversation comprises an opinion expressed or not, while the second step classifies the sentences expressing opinions as positive class, negative class or neutral classes. The research in OM in the mobile platform as it is in the infant stage.

1.1 Significance of Opinion Mining for Mobile Environment

Now-a-days, several large and small companies use OM in their business strategy. With the grand success and proliferation of mobile devices, the user's participation in social webs such as Twitter, FaceBook, and Whatsapp has increased tremendously. These data are used in several applications like advertising.OM is capable of providing useful information for businesses to improve product marketing, identify future enhancement of products, and manage the customer's feedback using various forms of expressions like reviews, ratings, and recommendations.

As the mobile users of social web generate a vast amount of unstructured data, it is important to mine the opinion.OM systems play a significant role in recommendation, detection of aggressive language in e-mail messages, improvement of the information extraction process [7] and summarization [8].Apart from this introduction part in section-1, rest of the paper is organized as: In section-2 related work is explained with different research papers on classification algorithms, sentiment classification across different domains. In section-3, a new problem statement is prepared to address the gap analysis. In section-4 we have explained the opinion mining frame work for social web .In section-5 we have pre-

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sented the result analysis with different machine learning classifiers for the proposed methodology. In section-6 conclusions of the work is discussed.

2. Related work

Movie rating and review-summarization system for a mobile platform is developed in [6]. SA determines the semantic orientation of the subjective terms. The movie rating data depends on the result of sentiment classification. The feature based summarization generates condensed the depiction of movie reviews. Latent Semantic Analysis (LSA) identifies product features. The system takes accuracy of sentiment-classification and response time into consideration.

The traditional machine learning algorithms such as Support Vector Machines (SVMs) and other classifiers [9] are used to perform SA on movie-review dataset, and outperform manually computed baseline algorithms [5].In accordance with the experimental results, SVMs performs better than the other classifiers and unigram with presence information proved to be the most useful feature.

Bootstrapping from a pair of two minimal groups of "seed" terms by calculating the number of hit results from the search engine with a NEAR operator determines the orientation of terms [3]. The result of the NEAR operator is successful only if two terms are within a specified word count of one another. The relationship between the given term and a group of seeds determines whether to classify under positive or negative subjectivity class.

The sentiment classification of user reviews across different domains is achieved using a sentiment sensitive thesaurus that lines up various phrases are expressing similar sentiment-details across several application-domains [10].It uses labeled data from different source domains and unlabeled data from both source and target domains for representing the "features" distribution. The primary step of the SA is to determine the semantic orientation of the terms in a sentence. A quantitative analysis of the glosses of subjective terms can be used to determine its orientation [4] in which glosses represent the definitions provided in the online dictionary. The terms with the same orientation are expected to have "similar" glosses. Therefore, use of synonyms and antonyms is capable of defining the orientation.

SENTIWORDNET [11] is a linguistic resource that assigns three scores such as Obj, Pos, and Neg for each WordNet synset. Obj, Pos, and Neg describe the objective, positive and negative nature of the terms in the synset. A method in [12] uses Twitter data to collect a corpus automatically for the OM. The sentiment variations on Twitter can be interpreted using LDA model with Foreground and Background LDA (FB-LDA) to extract foreground topics and background topics [13]. CISDL with n-fold cross validation which was discussed in our previous work [18] . The n-fold cross validation technique splits the data into n=10 folds and in each run it uses n-1 folds for training and n^{th} fold for testing. The process is repeated 'n' times and in each run the testing data is replaced with untested fold. The results obtained from each fold are then averaged to produce a single estimation.

3. Problem Statement

3.1 Gap Analysis

Due to several potential applications, the OM has gained increasing interest. Statistical approaches have been widely used for product feature selection to extract opinions. The features selected by these statistical approaches are sub-optimal due to its Non-Polynomial nature.

The existing machine learning based OM methods depend on labeled data from all domains to train a classifier and this may not adapt with a trained classifier in the absence of label information. Simple text categorization approaches could not mine opinions expressed in natural languages as these opinions are expressed in regional and language specific which are complex in nature. Sentiment classification in document level is not suitable for non-review based forum discussions, blogs, and news articles as these postings involve multiple entities.

3.2 Defining the Problem

OM in a mobile platform is challenging is task. The main problem of comments shared in the social network (social networking sites, blogs, micro-blogs, social book-marking and tagging) are often implicit and not explicitly stated by specific phrases that can be identified by phrases in lexical resources or by any other sources (seed lists).

Moreover, certain phrases and expressions are not even included in dictionaries like WordNet. Most of the comments and reviews that are shared via social-web in the mobile environment seems to be new phrases without necessarily having to refer to phrases previously used. The existing systems of sentiment classification find that a large number of features that is not feasible for the mobile environment. Most of the existing OM approaches involve only in observing the similarity between a phrase and a seed list of words that is not sufficient. Therefore, more sophisticated OM methods are required to solve the issues. The existing works covered only limited entities for opinion analysis.

It is not so easy to develop a mechanism or framework for identifying opinions with respect to a topic, as opinions are distributed all over the web especially in mobile view. There is some review presenting websites available such as Amazon.com and Yelp.com. These sites do not consider all entities and topics for sentiment classification. In this case, the task of opinion classification becomes formidable due to the explosion of diverse sites and the difficulty of classifying appropriate opinions. Though the techniques with unigram models achieve high accuracy, it is comparatively less than the topic based binary classification. Let us consider the Mobile view of customer reviews of apple mac book in Fig-1. In this review are posted with respect to a product along with rating, to ensure reliable and consistent user experience for consumers.



Fig-1. Mobile view of social media opinions

Reviews written by consumers provides the added advantage of displaying the seller's details, the product description which enables new buyers to take a quick decision while buying that product[17].

4. Opinion Mining Framework for Social Web

4.1 The new approach

Approach proposing here is an extension of the feature-based classification approach to identify the orientation of the phrase on

taking context into account to improve the accuracy. Natural language Processing deals with the syntax and semantics of languages could not make sense of exact opinion conveyed in a narrative manner .Therefore, the proposed methodology exploits natural language understanding systems, that translates the natural narrative information into formal form so that it is easier to interpret by a machine.

The OM can make use of these interpreted representations for further process. The lexicalized and statistical parsing followed by head parsing technique fine tunes the accuracy of opinion mining. It also focuses on syntax with semantics and long term relationship. Finally, the proposed work exploits a filtering mechanism to employ contextually dependent terms to reduce the opinion summary. Fig-2 depicts the framework of OM for mobile based social web.

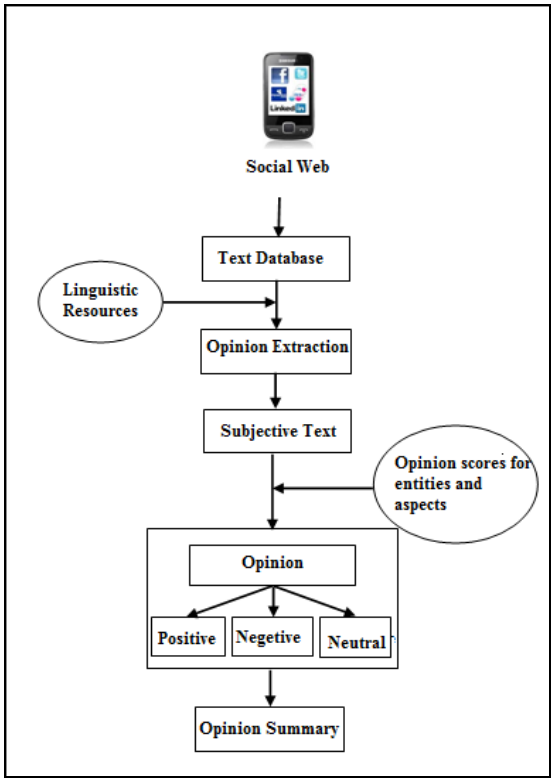


Fig 2: Opinion mining Framework for mobile based social web

4.2 Proposed Algorithm for the framework

S1. Design a framework for opinion mining for a mobile environment to assist mobile users in making decisions upon selecting a service or product (Retrieve mobile's social app data to crate text database step in Fig-2).

S2. Develop a novel scheme to identify the subject features appropriately to improve the accuracy and speed of opinion mining in a mobile environment (opinion extraction step in Fig-2). During this phase it also identifies negation phrases.

S3. Implement filtering mechanism called feature vector formation to let the users select the interested features for reducing the size of the summary of the opinions(filter the extracted opinions to form subjective step in Fig-2)

S4. Compute Opinion scores for entities and aspects from the subjective text (as shown in Fig-2) and then form feature vector.

S5. Perform Polarity categorization using feature vectors formed in step 4 both at sentence level and review level with Random Forest, NB and SVM algorithms.

S6. Compute performance analysis of the algorithms for the process with F-measure and ROC and repeat step 2 through step 5 until you achieved the desired level of accuracy.

5. Result Analysis

We have used hybrid supervised machine-learning technique for opinion extraction and obtained subjective text. A novel approach is used to filter and extract a required feature which is our subjective text. Since training data are labeled under two classes namely positive and negative for the sentence-level categorization, ROC (Receiver Operating Characteristic) curve is used to project the results depending upon the user perspective with different combinations of true positives, false positives, true negatives and false negatives for a better performance comparison. Performance of each classification model is estimated base on its averaged F-measure.

Where F-measure= (2xPr x R)/ (Pr+R). It is implemented for n number of classes where Pr is the precision of the selected class, R is the recall of the selected class, and n is the number of classes. Pri and Ri of each class of sample are evaluated using forest ,NB,SVM .Sentiment tokens and sentiment scores are information extracted from the original dataset, which will be used for sentiment categorization. In order to train the classifiers, each entry of training data needs to be transformed to a vector that contains those features, namely a feature vector.

An entity or a token is a word or a phrase that conveys opinion. In sentiment words of an ecommerce product review [15][17] ,a word entity consists of a positive (negative) word and its part-of-speech tag. In total, we have selected 5,478 word entities with each of them that occur at least 14 times throughout the review dataset. For phrase tokens, 1,065 phrases were selected of the 12,563, identified sentiment phrases, which each of the 2,012 phrases also has an occurrence that is no less than 30.

The statistical analysis of the word entities computed for the dataset is given in Table-1.

Table 1: Statistical information for word tokens

Entity	Mean	Median
Positive	3.21	3.15
Negative	2.22	2.6

5.1 Sentence-Level Categorization

We have generated the results on manually labeled sentences of mobile apps data (for testing purpose) and machine labeled sentences. With the help of the ROC curves plotted in Fig-3, it is clear to see that SVM, NB, RF classification algorithms performed quite well for testing data that have high posterior probability. As the probability getting lower, the NB classifier outperforms the SVM classifier, with a larger area under curve. In general, the Random Forest model performs the best.

With large set of feature vectors in the big dataset of live mobile apps with equal number of positive and negative labels are generated from the machine-labeled sentences, known as the complete set.

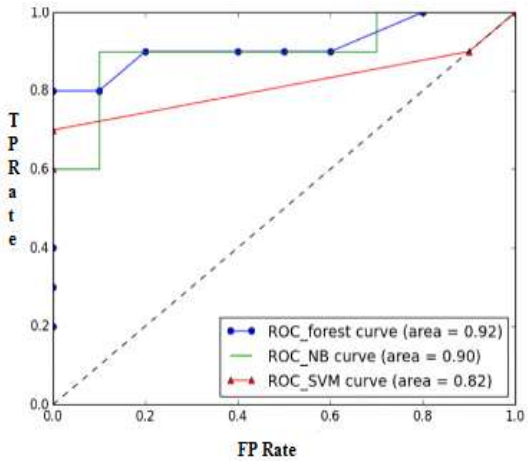


Fig 3: ROC with sample data set

Four subsets are obtained from the complete set, with subset .The amount of vectors with positive labels equals the amount of vectors with negative labels for every subset. Performance of the classification models is then evaluated based on five different vector sets (four subsets and one complete set, Fig- 4).

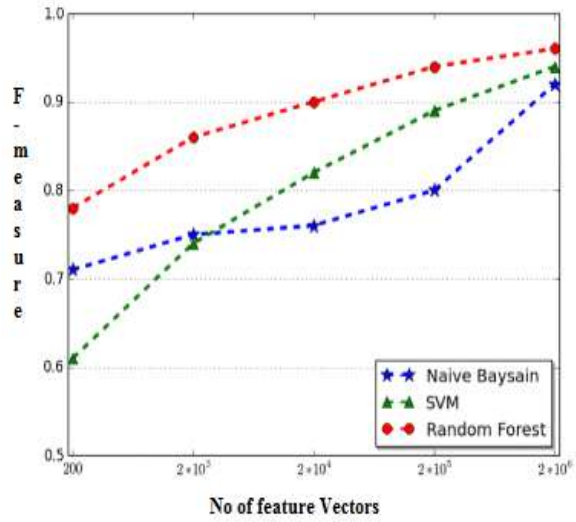


Fig 4: F-Measure for sentence-level categorization

In the curve plotted here, X-axis represents feature vectors count and Y-axis represents F-measure. With more training data, their measure is increasing. We can observe that NB classifier becomes the 2nd best classifier, on subset C and the full set. The Random Forest model again performs the best for datasets on all scopes. In Fig-5 we can observe that the ROC curves with TP rate as Y-axis and FP rate as X-axis plotted with complete data set.

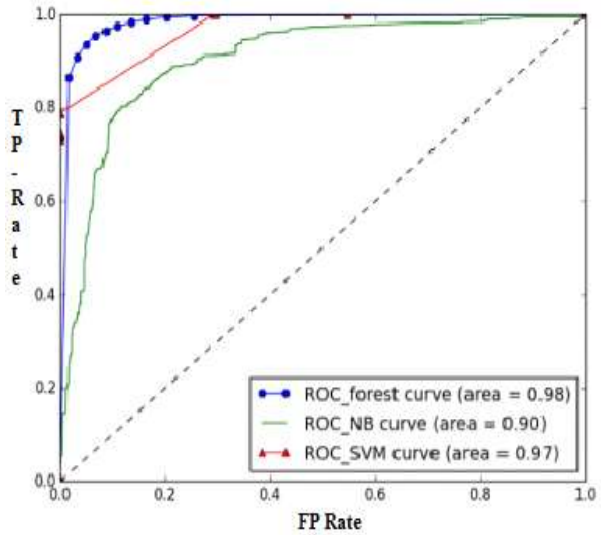


Fig 5: ROC on complete data set

5.2 Review-level Categorization

In Fig-6 the ROC curve shows the F-measure obtained on different sizes of vector sets. The curve clearly shows that both the SVM model and the NB model performances are identical. Both models are generally performs better than Random Forest model on all vector sets. However, neither of the models can reach the same level of performance when they are used for sentence-level categorization, due to their relative low performances on neutral class. In Fig-6 the ROC curve is plotted in which X-axis represents the number of feature vectors and Y-axis represents F-measure.

The experimental result is yielding good results, both in terms of the sentence-level categorization and the review-level categorization.

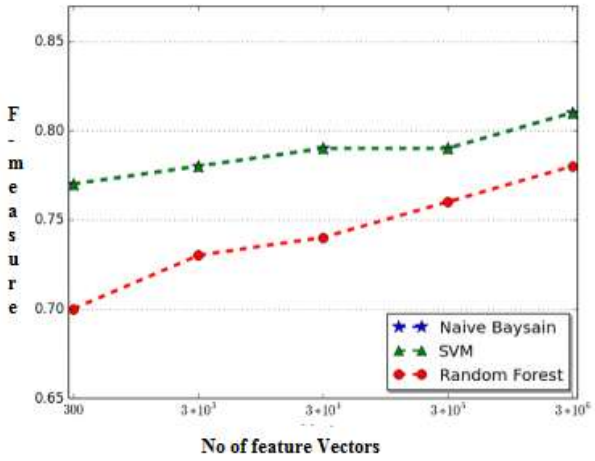


Fig 6: F-measure of review-level categorization

6. Conclusion

The mobile phone based social web activities become the most essential part of daily life. This data involves identification of opinion phrases and product features. This paper attempted to provides an Opinion classification with SVM, NB, RF and recommended an approach for OM for mobile based social web. In this paper we have analyzed different classification algorithms to classifying opinions both sentence level and review level and concluding that OM in a mobile environment is feasible only with a reduced number of features. F-measure and ROC are used to measure the performance of classification models. The proposed methodology achieves higher accuracy which allows the mobile users to make efficient decisions.

References

[1] Bing Liu, "Sentiment Analysis and Opinion Mining" Morgan & Claypool Publishers, pp. 1- 168, 2012

[2] Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis" Foundations and Trends in Information Retrieval, Vol. 2, No 1- 2, 1–135, 2008

[3] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews,"in Proceedings of 40th ACM Annual Meeting on Association for Computational Linguistics, pp. 417–424, 2002.

[4] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in Proceedings of 14th ACM International Conference on Information Knowledge Management, pp. 617- 624, 2005.

[5] B. Pang, L. Lee, and S.Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in Proceedings of ACM Conference on Empirical Methods in Natural Language Processing, Vol. 10, pp. 79–86, 2002.

[6] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, "Movie Rating and Review Summarization in Mobile Environment", IEEE transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 42, No. 3, pp. 397- 407, 2012

[7] Ellen Riloff, Janyce Wiebe, and William Phillips. "Exploiting subjectivity classification to improve information extraction" In Proceedings of 20th ACM national conference on Artificial Intelligence (AAAI), pp. 1106–1111, 2005.

[8] Yohei Seki, Koji Eguchi, Noriko Kando, and Masaki Aono. "Multi-document summarization with subjectivity analysis at DUC 2005". In Proceedings of the Document Understanding Conference (DUC), 2005.

[9] V. N. Vapnik, "The Nature of Statistical Learning Theory". Springer-Verlag, Information science and statistics, pp. 1- 314, 2000.

[10] Danushka Bollegala, David Weir, and John Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", IEEE transactions on Knowledge and Data Engineering, Vol. 25, No. 8, pp. 1719- 1731, 2013
[11] A. Esuli and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining," in proceedings of 5th Conference on Language Resources and Evalaution, pp. 417–422, 2006.
[12] Alexander Pak, and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" pp. 1320- 1326, 2010.
[13] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, an Xiaofei He, "Interpreting the Public Sentiment Variations on Twitter" IEEE transactions on Knowledge and Data Engineering, Vol. 26, No. 5, pp. 1158- 1170, 2014
[14] Chenghua Lin, Yulan He, Richard Everson, Member, IEEE, and Stefan Ruger, "Weakly Supervised Joint Sentiment-Topic Detection from Text" IEEE transactions on Knowledge and Data Engineering, Vol. 24, No. 6, pp. 1134- 1145, 2012
[15] Amazon customer - reviews , http://www.amazon.in/gp/cdp/member-rviews/A3A9YE4SZADQ5W/ref=cm_cr_tr_tbl_1_sar?ie=UTF8&sort_by=MostRecentReview
[16] Jiang, Long, et al. "Target-dependent twitter sentiment classification."Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.
[17] EBay product review in mobile app,http://pages.ebay.com/sellerinformation/news/sprupd16/product-reviews.html
[18] Salina Adinarayana, E.Ilavarasan, "An Efficient Decision Tree for Imbalance data learning using Confiscate and Substitute Technique.", Materials Today: Proceedings , pp. 680-687,2018 Volume 5,Issue 1P1,ISSN 2214-7853