Survey on Lifelong Learning for Large-Scale Social Media Sentiment Analysis

Kumud Shende
Department of Computer Engineering
Pune Institute of Computer Technology, Pune, India

Prof. Kalyani Waghmare
Department of Computer Engineering
Pune Institute of Computer Technology, Pune, India

Abstract: This survey paper review based on the Lifelong learning sentiment analysis for social media texts. Lifelong learning means the people continuous learning process or lifelong learning aims to learn as peoples do: retain the learned knowledge from the previous task and use it to help in the future learning task. In the social media that contain a large range amount of text and a large range of topics, so it would be very difficult to manually collect enough labeled data to train a different sentiment classifier for different domains. In social media, the text is continuously increasing and constantly changing the topics. So we only focused on large scale data-sets and sentiment analysis different technique. Now a day users are relying on social media, so the importance of a review is going higher. Sentiment analytic thinking plays a classifying increasingly more important role in the user's opinion, attitude, and expressed their feeling in a text, so we focused sentiment analysis. But in machine learning, going through one thousand text reviews would be much easier, if any model is used to polarize those reviews and learn from it. We used various algorithms in the literature survey like Naive Bayes's, support vector machine and Maximum Entropy for the lifelong long learning sentiment analysis. In this paper, a brief survey is carried out on various data mining and machine learning algorithms for Lifelong learning for Large-Scale Social Media Sentiment Analysis. A method on large scale data-sets to polarize different classification like positive or negative and gives better accuracy.

Keywords— *Sentiment analysis, lifelong learning, social media analysis, NLP*

I. INTRODUCTION

Sentiment Analysis is the task of classifying the different categories like positive and negative feeling. Sentiment analysis uses of the natural language processing, machine learning, text analysis, and computational techniques automate the classification of sentiment analysis from sentiment review. Analysis of these opinions and sentiment has spread many other fields such as customer information, different websites books, marketing, and socials. In daily life, a thousand of people depend on the online sentiment rating, review or comments. 90% of the users' decision depended on the online rating and review. Sentiment analysis is a classification of the give different text polarity at these levels such as document level, Aspect level, and sentiment level. Social media allow millions of users to express their feelings and freely spread their opinions about the particular related topic and also show their attitudes by liking or disliking content. Sentiment refers to the user's emotions or opinion about different event, entities, and ideas. In now a day social media generate high volume, velocity, variety, variability data on social media. In social media data like Twitter data are 80% of data is based on the text, therefore text classification has become very important for public sentiment and opinion elicitation. Then it consists of classifying the different posts' polarity such as positive and negative. Machine learning technique and topic are depended on lifelong learning, transfer learning, never-ending learning, multi-task learning, self-taught learning and online learning, no unified definition available for lifelong learning. A lifelong learning system needs some general components: Past Information Store (PIS), Knowledge Base (KB), Knowledge Miner(KM), and Knowledge-Based Leaner.

II.Literature Review

The author in [1] proposed a system for lifelong sentiment learning that uses machine learning algorithms. In this paper, the author introduces a novel approach for automatically classifying the different sentiment of Twitter messages. These messages are classified into different polarity such as positive or negative sentiment. This classifier is useful for a user who wants to search for the new product before purchase. The author has used three different algorithms, such as Naive Baye's, SVM and Maximum Entropy. In this paper, the author has used performance evaluation measure. Among all the three algorithms Naive Baye's gives good accuracy. In This paper are used the two data sets in the languages of both English Twitter and Chinese Weibo , and they are testing it on nine benchmark sentiment analysis data-sets. In this paper, the author proves that our lifelong sentiment learning approaches are feasible and effective to tackle the different social media related challenges and also prove that space," we take more training data that give the best performance", so it does not hold in large scale social media sentiment analysis. In this paper, the author mentioned some challenges have arisen in sentiment analysis. 1)The large scale amount of text data in social media is massive and continuously increasing data and latest topics in social media are continuously changing.

In [2] the author introduces to approach for automatically classifying the various sentiments of Twitter messages. These Twitter messages are classified into different categories such as positive or negative feelings. Twitter messages are used for people who want to search the review of products before purchase, and a lot of companies that want to see the user sentiment of their product or brands. In this paper, there is no previous research based on classifying Twitter messages on micro-blogging services like Facebook, Twitter etc. Machine learning algorithm used Naive Baye's, Maximum Entropy, SVM gives better accuracy is above 80% when trained with emoticon data. Machine Learning gives reliable output over the input provided by the user. This paper describes the preprocessing steps needed in order to achieve better accuracy. In this paper, the researcher does not consider neutral tweets have been our training or testing data-set. They only use positive or negative tweets. The researcher is to use different machine learning classifier and feature extraction. The feature extraction is based on bi-gram, uni-gram, with part of speech tags. The researcher using noisy labels data for training. In this paper, the experiment was done on a movie review.

The author in [3] proposed zero short learning and one short learning, which only used in a small number of examples or even no example to learn. However, for lifelong machine learning technique were proposed in the context of memory-based learning and neural network. The lifelong learning which considers that can learn various tasks from one or more domain. This one or more domain is used in a lifetime. Lifelong machine learning is a machine learning that learns human continuously learning the process. The main goal is lifelong learning to retain the knowledge learned from past task and used it then help with used future learning.

The work cited by [4] author gives details about sentiment analysis used natural language processing. This review paper mentioned the recent development work in natural language processing research to look at the past, present, and future of the natural language processing technology. The "jumping curves" from the field of marketing prediction and business management. In natural language processing, the researcher has been focusing on tasks such as information retrieval, question answering, machine transaction, text summarization, and opinion mining. natural language processing research focuses on syntax because syntactic processing was manifestly necessary.

In [5] this paper researcher gives details about sentiment analysis in social media text has been studied in depth. The author gives a full detailed about the different emotions. Emotion plays an important role in daily life, human communication. Emotional intelligence is more important than IQ. An Effective computing and sentiment analysis have great potential as a sub-

component of their system. They can enhance the great capabilities and other recommendation systems. The important basic task of Effective computing and sentiment analysis. Effective computing and sentiment analysis are important in polarity detection and emotion recognition. The sentiment analysis and Affective Computing are based on the polarity detection, emotion recognition and multi-modal fusion. In this paper two challenges have arise, such as opinion target identification and subjectivity detection. Example. The author mentioned in different polarity detection such as positive versus negative, thumbs up v/s thumbs down, etc. example Twitter messages can be classified as good or bad news without being subjective. The existing approaches to sentiment analysis and effective computing into three categories: statistical methods, hybrid approaches and knowledge-based technique. This task is integrated and interdependent to the sentiment categorization model.

Author on [6] paper use the different feature for detecting the sentiment of Twitter messages. To evaluate the existing lexical information about the informal and creative languages used in micro-blogging. The leverage existing hash tags in the twitter large scale data for a building training data-set. The sentiment lexicon has proved that the sentiment analysis is useful for other domain and it will they also prove useful for sentiment analysis in twitter. In this paper, the author explores some method for building to date using Twitter hashtags to identify the positive, negative and neutral tweets to use for training the three sentiment classifier. In this paper, the author uses three different corpus of Twitter messages in our experiment. For training, we use the hash-tagged data set (HASH), which compiler for the twitter corpus and emotion data-set (EMOT) used for the evaluation annotated dataset produced by the iSieve corpus an (ISIEVE). Our experiment on Twitter sentiment analysis shows that part of the speech feature may not be useful for sentiment analysis in the other domain. Using the hashtag to collect training data did prove useful, and did use the data collected based on positive and negative emotions. This experiment shows that the micro-blogging feature is included.

The work cited by [7] the author use previously proposed state-of-the-art uni-gram models as our baseline and report an overall gain of over 4 % for two classifications. In this paper use a 3-way classification task with different classes of sentiment polarity such as positive, negative and neutral. Author use a balanced data-set of 1709 instances to each class and baseline is 33.33%. In this paper author has used different performance evaluation measures such as F-Measure, accuracy and Learning curve. The investigate the two models like Tree kernel and feature-based models and demonstrate both models perform the uni-gram baseline. For future based approach, this feature is that combine the polarity of words and their part-of-speech-tags. Sentiment analysis of Twitter data is not different from sentiment analysis for other genres.

The author in [8] proposed the use of semantic feature in Twitter sentiment classification and explores three different approaches like replacement, augmentation, and interpolation. Author has used classification algorithms such as Naive Baye's. In this paper compare the performance of semantic sentiment analysis approaches against the baselines like Uni-gram Feature used for sentiment analysis of tweets data and Part-of-speech-Feature used in literature for the task of Twitter sentiment analysis. Used STC, HCR and OMD data-set. In this paper author has used different performance evaluation measures to compare the Precision, Recall and F-measure of our semantic sentiment analysis against the baseline. Semantic feature produces higher Recall and F1 score , but lower in precision, so then sentiment feature when classifying negative sentiment and also show the semantic feature outperforms the sentiment feature for positive sentiment classification in precision , but not in Recall and F1 score.

The author in [9] the researcher studied in depth on target-dependent Twitter sentiment

classification by use of rich text automatic features based on word representation. In this paper has used Support vector machine algorithm and used various model such as Target-ind , Target-dep-, Target-dep, Target-dev+ on Twitter .Previous work relies on syntax, such as automatic sparse trees. In this paper, the author shows the competitive result can be achieved without the use of syntax, by extracting rich text set of automatic features. Our experiment shows that multiple pooling functions, multiple embeds and sentiment lexicons used the rich text of feature information, which improved the accuracy.

The author in [10] proposed a three-stage model, namely Polarity Shift Detection, Elimination and Ensemble (PSDEE) to address the polarity shift problem in the document level sentiment analysis. The author proposed a hybrid model that employs both statistics based and rule-based methods to detect different type of polarity shift. Firstly proposed a hybrid polarity detection approach, which is a rule-based method to detect the polarity shifts such as explicit negation, contracts, and inconsistencies. Secondly, polarity shift elimination algorithm to eliminate polarity shift in negation . This paper has used the pseudo-code of the hybrid polarity shift detection algorithm, linear SVM, logistic regression and Naive Baye's. For naive Baye's, use the OpenPR_NB toolkit, Logistic regression used the LibLinear toolkit, linear SVM used the LibSVM toolkit. In this paper has used Multi-domain data set. It consists of domains such as Book, Kitchen appliance, Electronics, and DVD of review from Amazon.com. Each of the four data-sets contains 1000 positive reviews and 1000 negative reviews. The polarity shift is a major factor that affected the classification performance of machine learning and sentiment analysis system. Their results demonstrate the effect of our PSDEE approach compared to several related works that address a polarity shift in document-level sentiment classification.

The author in [11], the proposed method can utilize sentiment relation between messages to facilitate sentiment classification and effectively noisy data. An author investigates whether social reactions can help sentiment analysis by proposing a sociological approach to handling short text (SANT) and noisy data for sentiment classification. In this paper used optimization algorithm for SANT. This paper used a mathematical model optimization formulation that used in the sentiment considering and emotional contagion theories into the supervised learning process, and also utilize the sparse learning to tackle the noise and mess texts data in micro-blogging. In this paper used the two publicly available twitter data-sets are employed such as standard twitter sentiment (STC) and Obama-McCain Debate (OMD) data-set. In the paper author approach novel sociological approach (SANT) to handled networked texts in micro blogging posts. In this paper, the author has used different performance evaluation measures such as Sensitivity, Specificity, F-Measure, Precision and graph Laplacian. The experiment result shows the user-centric social relation is helpful for sentiment classifier of micro-blogging messages.

The [12] author has used different classification algorithms such as Spam Filter algorithm and Spell Checker Algorithm. An introduced a Twitter-based sentiment analysis system and also based on the different topic searched, TwiSent collects the tweets pertaining to it and categories them into three polarities. In this paper analyzing the micro-blog posts has faced many challenges as compared to other text genres like News, Blogs. In this paper, tackle the some problems which are 1) Twitter-based spam 2) Spell checker for noisy text and structural anomalies in the text in the form of incorrect spellings, nonstandard abbreviations, slangs, etc. 3) Entity detection in the context of the topic searched and 4) Pragmatics embedded in the text. For overall system, perform a 2-class and a 3-class classification using TwiSent. In the 2-class classification considers only positive and negative

tweets. In the 3-class classification considers positive, negative and all objective tweets. In this paper author has used different performance evaluation measures such as Sensitivity, Specificity, F-Measure, Precision. In this paper not only mentioned the issues with the microbiology but also present an effective system to handle them. An Author shows the system performance much better than an existing system. Also, show that a system performance of an auto-annotated data-set does not guarantee similar performance on real-life micro-blog data.

III. TABLE

TABLE I. OVERVIEW OF LIFELONG LEARNING SENTIMENT ANALYSIS

Sr. No.	Author	Objective	Techniques Used	Dataset	Limitation	Challenges	Accuracy
1]	Doaa et al.(2015)	Bag of word model solving the challenges of sentiment analysis.	Bag of words model are used.	1000 training set, 5000 test set, 10.000 verified set.	Less accuracy in BOW , neglect grammar.	Lexicon, feature extraction, negation, word knowledge	SAOOP 83.5% that better than the BOW 62%.
2]	Chetan and atul(2014)	Lexicon based technique for sentiment analysis using natural processing languages and machine learning.	Lexicon based techniques are used.	560 Chinese review dataset.	Fail to efficiently handle the vast amount of sentiment data.	Huge lexicon	73.5%
3]	Svetlana et al.(2014)	To detect the sentiment of short informal textual messages such as tweets and SMS.	SemEval-2013	2000 pos words &47000 neg words.	Limited in length.	Domain dependence, short informal textual messages, detecting the sentiment a state-of-art sentiment analysis.	Improve accuracy and F- score of 69.02% on the tweets test and 68.46 % on the SMS test set.
4]	Shoushan Li et al. (2013)	Active learning for cross-domain sentiment classification	Bag of Words	Book, DVD, electronics, kitchen appliance	Limited size of the labeled data in the target domain.	Domain dependence, Domain adaption problem.	81.5%, 80.5%,83.8%,86.5%

		by actively selecting a small amount of labeled data in the target domain		review, 1000 pos &1000 neg.			
5]	Andrius et al.(2012)	A concept level sentiment analysis system that seamlessly integrates into opinion mining lexicon based and learning-based approaches.	Bag of words, Support Vector machine.	Software review, Movie review	Limited information about the sentiment topic or rationale .	Huge lexicon.	82.30%
6]	Rui Xia el al. (2011)	Focus on the tasks cross-domain sentiment classification	Part-of-Speech.	1000 pos &1000 neg review, Multi-domain sentiment data-set.	Nouns become less important.	Domain dependence	Uni baseline by 3.01% &3.94%, 0.93% higher
7]	Yasuhisa el al. (2011)	Multiple domain sentiment analysis identifying domain dependence and independence world polarity.	Part-of-Speech.	Used 17 domains &1000 dic from the multi-domain sentiment	Cannot handle multiple. source domains &multi-target domains	Domain Dependence , multi domain dependence	Not present in the baseline model.
8]	Bas et al.(2011)	Investigate the impact of accounting for negation in sentiment analysis.	Part-of-speech.	Dutch languages, 13,628 human rated dutch document on 40 different topic.	Explored only to a limited extend.	Negation	71.23 % for negation, precision improves with 1.17% from &0.41% without negation.
9]	Yulan et. al(2011)	Domain Adaptive using joint sentiment topic model	Naive Bayes, SVM.	Movie review, Multi domain data-set.	To detect sentiment and topic simultaneously from text	Domain dependence	90%
10]	Maral(2011)	The negation detection in sentiment evaluation using BOW	BOW term frequencies.	2000 movie review : 1000 positive	No significant different in classification accuracy when different	Negation and Domain dependence	Sentiment analysis without negation was 66% &positive rate(recall) was 84% &precision

		using term frequencies to evaluate the discrimination capacity of our system with different window size.		&1000 negative.	window sizes have been applied.		was 62%.
11]	Erik and Francine(2009)	Sentiment analysis in multilingual web texts using machine learning.	NLP, machine learning, information retrieval.	Blog, review dataset.	Limited no. Of annotated training, manual labeling is limited.	NLP overheads (Multilingual), opinion extraction from noisy web texts (such as blog) still poses problem.	83%, 70% and68%.

IV. CHALLENGES OF SENTIMENT ANALYSIS

1) Many tweets do not have feelings, so it is a current limitation of our research to not include the neural network.

2) A short length of a status update: The character of a social media text such as short length, large scale, and dynamic topic has brought new challenges to the research of sentiment analysis.

3) Open Domain Dependency: The polarity shifted problem is a big problem that effected in the classification performance in machine learning and sentiment analysis. The polarity changes from the one domain to another domain in domain dependency. Sentiment shifter in is based on word and phrases that include negation word and contracts and etc. In negation, 1)ex. I don't like the blue dress. The negative word is "don't" this word shifted the polarity of the sentiment word "like".2)Contracts: ex Farhan Akhtar doing fairly good acting," but" shifted sentiment polarity previous phrases "fairly good acting ".

4) Open Domain Dependency: English languages are mostly used because people know the English languages with other languages than English like Marathi, Hindi, etc. I.e lexicons dictionaries for these languages are different.

5) Fake tweets opinion about a user: Sentiment people give the fake review about any product, movie, etc. The fake opinion is misguided the users by providing them into untruthful opinion related to the positive or negative sentiment.

6) Negation polarity: In sentiment analysis challenging task is negation and the negation is usually changes the opinion polarity.

METHODOLOGY

Method of Sentiment Analysis:

• **Data Collection:** Opinion and feeling are expressed in a different way, with a different vocabulary, the context of writing, usage of short forms and large scale of data, making the data is huge and disorganized. Manual analysis of sentiment data is almost impossible. Therefore programming languages like R languages are used to process and analyze the data.

• **Text Preparation**: Text preparation is filtering and that extracted data before analysis.

• **Sentiment Detection:** At this stage, each sentence of the review of public and opinion is examined for subjectivity. Sentence with subjective expression are retained that which convey objective expression are discarded. Sentiment analysis is done at different levels using common computational techniques like uni-gram, lemma, and negation and so on.

•**Sentiment Classification:** At the stage sentiment analysis methodology, each subjective sentence

detected is classified into groups positive, negative, good, bad like dislike. Presentation Output: The main idea of sentiment analysis of data is to convert unstructured text into meaningful information.

Conclusion

In this paper a brief review is carried out on different data mining and machine learning algorithms for lifelong learning for Large-Scale Social Media Sentiment Analysis. Various technique involve in phrases and word classification method. Various machine learning technique have been used in the paper mentioned. In this survey paper the most common algorithms are Naive Baye's, Support vector machine and Maximum Entropy. The machine algorithms depend upon the various feature extraction and this feature have been used in Uni-gram, part-of-speech, bi-gram, emotion detection etc. Different algorithms give different accuracy with different number of parameters for the prediction.

REFERENCES

[1] Xia, R., Jiang, J., & He, H. (2017). Distantly supervised lifelong learning for large-scale social media sentiment analysis. IEEE Transactions on Affective Computing, 8(4), 480-491.

[2] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12), 2009.

[3] Chen, Z., & Liu, B. (2016). Lifelong machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 10(3), 1-145.

[4] Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine, 9(2), 48-57.

[5] Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2), 102-107.

[6] Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In Fifth International AAAI conference on weblogs and social media.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in Proc. Workshop Languages Social Media, 2011, pp. 30–38.

[8]Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis of twitter. In International semantic web conference (pp. 508-524). Springer, Berlin, Heidelberg.

[9] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 151-160). Association for Computational Linguistics.

[10] Xia, R., Xu, F., Yu, J., Qi, Y., & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. Information Processing & Management, 52(1), 36-45.

[11] Hu, X., Tang, L., Tang, J., & Liu, H. (2013, February). Exploiting social relations for sentiment analysis in microblogging. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 537-546). ACM.

[12] Mukherjee, S., Malu, A., AR, B., & Bhattacharyya, P. (2012, October). TwiSent: a multistage system for analyzing sentiment in twitter. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2531-2534). ACM.

[13] El-Din, D. M., Mokhtar, H. M., & Ismael, O. (2015). Online paper review analysis. International Journal of Advanced Computer Science and Applications (IJACSA), 6(9).

[14] Kaushik, C., & Mishra, A. (2014). A scalable, lexicon based technique for sentiment analysis. arXiv preprint arXiv:1410.2265.

[15] Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research, 50, 723-762.

[16] Li, S., Xue, Y., Wang, Z., & Zhou, G. (2013, June). Active learning for cross-domain sentiment classification. In Twenty-Third International Joint Conference on Artificial Intelligence.

[17] Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining (p. 5). ACM.

[18] Xia, R., & Zong, C. (2011). A POS-based ensemble model for cross-domain sentiment classification. In Proceedings of 5th International Joint Conference on Natural Language Processing (pp. 614-622).

[19] Yoshida, Y., Hirao, T., Iwata, T., Nagata, M., & Matsumoto, Y. (2011, August). Transfer learning for multiple-domain sentiment analysis identifying domain dependent/independent word polarity. In Twenty-Fifth AAAI Conference on Artificial Intelligence.

[20] Heerschop, B., van Iterson, P., Hogenboom, A., Frasincar, F., & Kaymak, U. (2011). Accounting for negation in sentiment analysis. In 11th Dutch-Belgian Information Retrieval Workshop (DIR 2011) (pp. 38-39). [21] He, Y., Lin, C., & Alani, H. (2011, June). Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 123131). Association for Computational Linguistics.

[22] Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval, 12(5), 526-558.

.