National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis

AAMERA Z.H.KHAN **Dr. MOHAMMAD ATIQUE** **Dr. V. M. THAKARE**

***Abstract* — Sentiment analysis is a growing area of research with significant applications in both industry and academia. Most of proposed solutions centered around supervised, and machine learning approaches. Twitter's unique characteristics give rise to new problems for current sentiment analysis methods, which originally focused on large opinionated corpora such as product reviews. This paper presents a new entity-level sentiment analysis method for Twitter. The method first adopts a lexicon based approach to perform entity-level sentiment analysis. This method can give high precision, but low recall. To improve recall, additional tweets that are likely to be opinionated are identified automatically by exploiting the information in the result of the lexicon-based method. A classifier is then trained to assign polarities to the entities in the newly identified tweets.**

***Key Words* — Sentiment analysis, machine learning, twitter.**

I. INTRODUCTION

Sentiment analysis or opinion mining is an important type of text analysis that aims to support decision making by extracting and analyzing opinion oriented text, identifying positive and negative opinions, and measuring how positively or negatively an entity is regarded[1,6]. An increasing number of people are willing to post their opinions on Twitter, which is now considered a valuable online source for opinions. As a result, sentiment analysis on Twitter is a rapid and effective way of gauging public opinion for business market and social studies. A tweet is a text-based post and only has 140 characters, which is approximately the length of a typical newspaper headline and subhead. The short messages are very easy and convenience to both sender and reader to share things of interest and communicate their thoughts anywhere and anytime in the world. Twitter is a "what's – happening-right-now" social network and hence offer immediate sentiment. The amount of digital information available to individuals is ever-increasing. A considerable amount of this information is in textual format, which can be broadly categorized into two main types: Facts and opinions [2, 7]. Facts indicate objective information whereas opinions can be subjective and indicate the sentiment of the author about an issue. Opinions can be about anything, e.g. a product such as a movie, a service such as food service of a restaurant or a company such as Amazon. Sentiment analysis, the process of automatically detecting if a text segment contains emotional or opinionated content and determining its polarity, is a field of research that has received significant attention in recent years, both in industry and academia [3, 8].It is an exciting new research field with the potential for a number real world applications where discovered opinion information can be used to help people or companies or organizations to make better decisions. As a micro blogging and social networking website, Twitter has become very popular and has grown rapidly. Twitter is an ideal source for spotting the information about societal interest and general people's opinions. However, there has been little prior opinion mining work in the micro blogging area since Twitter is a relatively new technology.

With the rapid development of internet distributed computing up to now, it is easy to analyze tremendous amount of data and predict customer favorites and future demand [4]. Sentiment analysis has come to play a key role in text mining applications for customer relationship management, brand and product positioning and market research. In spite of recent advances, there are still several promising new directions for developing and advancing sentimental analysis research [5].

II. BACKGROUND

There are many researches happen in the area of sentiment analysis. To determine whether a document or a sentence expresses a positive or negative sentiment, two main approaches are commonly used: the lexicon-based approach and the machine learning-based approach. The lexicon-based approach determines the sentiment or polarity of opinion via some function of opinion words in the document or the sentence. The machine learning-based approach typically trains sentiment classifiers using features such as unigrams or bigrams. Most techniques use some form of supervised learning by applying different learning techniques such as Naive Bayes, Maximum Entropy and Support Vector Machines. These methods need manual tagging of training examples for each application domain. While most sentiment analysis methods were proposed for large opinionated documents (e.g. reviews, blogs), some recent work has addressed micro blogs. Supervised learning is the dominant approach. Many Twitter characteristics and language conventions (e.g. hash tags and smiley) were utilized as features. There are also several online Twitter sentiment analysis systems. These approaches mainly used supervised learning.

III. PREVIOUS WORKDONE

As more and more users express their political and religious views on Twitter, tweets become valuable sources of people's opinions. Xujuan Zhou et.al. [1] introduced a Tweets Sentiment Analysis Model (TSAM) that can spot the societal interest and general people's opinions in regard to a social event. The TSAM model will yield much more accurate results by taking a number of research issues into consideration including distinguishing between parts of speech, taking emotion analysis into account and utilizing more accurate entity recognition techniques. The application of sentiment analysis on extracting the quality attributes of a software product based on the opinions of end user that have been stated in micro blogs such as Twitter presented by Rahim Dehkharghani et al. [2].This technique gives advantageous such as document frequency of words in a large number of tweets. The extracted results can help software developers know the advantages and disadvantages of their products. GEORGIOS PALTOGLOU et al. [3] an intuitive, less domain-specific, unsupervised, lexicon-based approach that estimates the level of emotional intensity contained in text in order to make a prediction. An approach can be applied to, and is tested in, two different but complementary contexts: subjectivity detection and polarity classification. The results demonstrate that the proposed algorithm, even though unsupervised, outperforms machine learning solutions in the majority of cases, overall presenting a very robust and reliable solution for sentiment analysis of informal communication on the Web.

In [4] Zhang et al. concluded unique characteristics through more than 1,400,000 real mobile application review (1) Short average length (2) Large span of length (3) Power-law distribution and (4) Significant difference in polarity. Another model called Combined Sentiment Topic (CST) model to detect sentiments and topic simultaneously from text discovered by M.S.Usha et al [5]. This model is based on Gibbs sampling algorithm. Besides, unlike supervised approaches to opinion mining which often fail to produce good performance when shifting to other domains, the unsupervised nature of CST makes it highly portable to other domains. CST model performs better compared to existing semi-supervised approaches.

IV. EXISTING METHODOLOGY

The Tweets Sentimental Analysis Model (TSAM) [1] automatically analyses tweets data. It can identify the positive, negative or neural opinions and measure intensity of positive or negative opinions in regard to entity. The conceptual framework of the TSAM consists of three modules: Feature selection module that extracts the opinionated words from each sentence. Sentiment identification module that associates expressed opinions with each relevant entity in each sentence level. Sentiment aggregation and scoring module are the terms that are used to calculate the sentiment scores for each entity. Machine-Learning approaches [3] were among used to explore the sentiment analysis of reviews. Their best accuracy attained in a dataset consisting of a movie reviews used a SVM classifier with binary features. Most other approaches in the field have focused on extending the feature set in order to improve the accuracy. The lexicon based classifiers a typical example of an unsupervised approach, because it can function without any reference corpus and does not require any instruction .The classifier is based on estimating the intensity of negative and positive emotion in text in order to make a ternary prediction for subjectivity and polarity, that is, the output of the classifier is one of {0, +1, −1}.A document is classified as objective if its scores are {+1,-1}.All dictionary lemmas are stemmed using the porter stemmer.

V. ANALYSIS AND DISCUSSIONS

One nontrivial task tweets data collection for sentiment analysis is the extraction of the relevant entities from the tweets. System performance is measured with its accuracy as the ratio of NTSCL to TNTTS where NTSCL is the number of tweets the system correctly labeled and TNTTS is the total number of tweets in a test set [1].Results from subjectivity detection show that the lexicon based classifier is able to perform very adequately and offers a reliable solution. Building a lexicon-based sentiment intelligent system is doable and can be very beneficial. In [3] the problem of sentiment analysis in social networking Media addressed, such as MySpace, Twitter, Digg, forums, blogs, etc. argued that this area of application provides unique challenges, not addressed in typical review-focused sentiment analysis environments. An intuitive, unsupervised, lexicon-based algorithm which estimates the level of emotional strength in text in order to make a final prediction. Proposed solution is applicable to two complimentary tasks: subjectivity detection and polarity classification, overall providing a comprehensive solution to the problem of sentiment analysis of informal communication on the Web. The advantages of the approach are that it requires no training and thus can be readily applied into a wide selection of environments.

VI. PROPOSED METHODOLOGY

Figure 1 gives an architectural overview of sentiment analysis algorithm.

Twitter Data

Twitter has developed its own language conventions.

Preprocessing

Before starting sentiment analysis, need to do some data cleansing. After cleaning, perform sentence segmentation, which separates a tweet into individual sentences. Afterwards, tokenize and perform part of speech tagging (POS) for each sentence.

Augmented lexicon-based method

In this section, propose an augmented lexicon based approach to sentiment analysis considering the characteristics of the Twitter data.

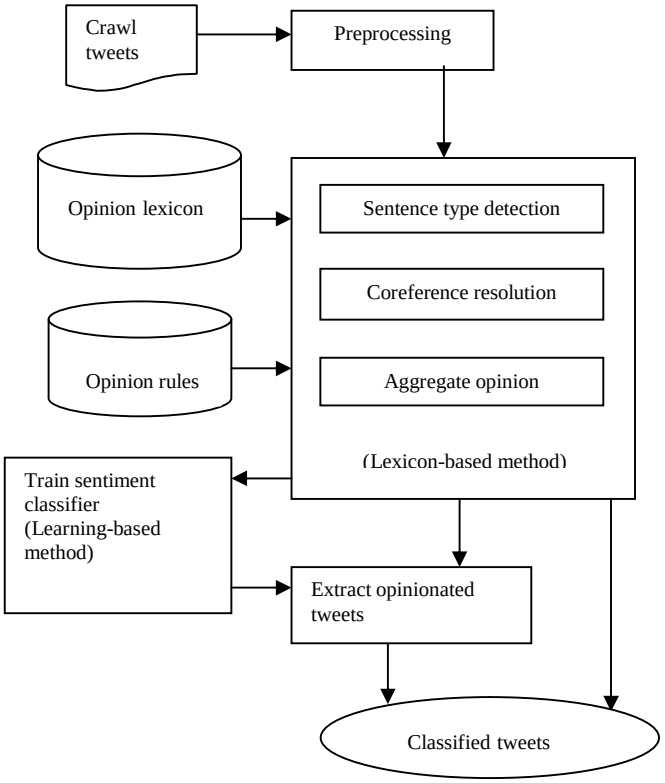

Fig. 1: Algorithm architectural overview

Decreasing and increasing rules: This set of rules says that deceasing or increasing the quantities associated with some opinionated items may change the orientations of the opinions. For example, "The drug eases my pain greatly". Here "pain" is a negative opinion word in the opinion lexicon, "and the reduction of pain" indicates a desirable effect of the drug.

VII. Possible Outcome And Result

First use accuracy to evaluate the whole classification performance of the method with three classes, positive, negative and neutral. For positive and negative sentiments on entities, employ the standard evaluation measures of precision, recall and F-score. The method for identifying sentiment indicators can get many sentiment orientations wrong, which causes mistakes for the subsequent step of sentiment identification using the lexicon-based method. The proposed method outperforms all the baseline methods by large margins in identifying opinions on entities.

Conclusion

The unique characteristics of Twitter data pose new problems for current lexicon-based and learning-based sentiment analysis approaches. In this paper, proposed a novel method to deal with the problems. An augmented lexicon-based method specific to the Twitter data was first applied to perform sentiment analysis. Through Chi-square test on its output, additional opinionated tweets could be identified. A binary sentiment classifier is then trained to assign sentiment polarities to the newly-identified opinionated tweets, whose training data is provided by the lexicon-based method. In future, the goal is to extend the applicability of the proposed solution to other languages, for which training data is especially difficult to come by.

References

[1]Xujuan Zhou , Xiaohui Tao, Jianming Yong ,Zhenyu Yang," Sentiment Analysis on Tweets for Social Events" ,*proceedings of IEEE*, pp-557-561, 2013.
[2] Rahim Dehkharghani and Cemal Yilmaz, "Automatically Identifying a Software Product's Quality Attributes through Sentiment Analysis of Tweets" ,*IEEE*, pp-25-30 ,2013.
[3] GEORGIOS PALTOGLOU and MIKE THELWELL "Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media", *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4,pp-66-66:19, September 2012.
[4] Lin Zhang, Kun Hua , Honggang Wang ,Guanqun Qian ,Li Zhang, "Sentiment Analysis on Reviews of Mobile Users" ,*Science direct*,pp-458-465, 2014.
[5] M.S.Usha, Dr.M.Indra, "Analysis of Sentiments using Unsupervised Learning Techniques".
[6]S.Brody and N.Elhadad ,"An unsupervised aspect-sentiment model for online reviews", The Annual Conference of the North American in Human Language Technologies(HLT),pp 804-812 ,2010.
[7]A.Hamouda and A.Rohaim,"Reviews classification using sentiwordnet lexicon ",*in Jounal on Computer Science and Information Technology(OJCSIT)*,vol. 2,no.1,201.
[8]BACCIANELLA, S. ESULI,A.,AND FABRIZIO,S, "Sentiwordnet 3.0:An enhanced lexical resource for sentiment analysis and opinion mining", *proceeding of the Annual Conference on Language Resource and Evaluation(LREC).*

Author's Profile

Aamera Z.H.Khan
M. E. pursuing, Department of Computer Science and Engineering SGBAU, Amravati. Email.ID: khan.aamera01@gmail.com

Dr. Mohammad Atique
Dr. Mohammad Atique is presently working as Associate Professor in Computer Science & Engineering in PG Teaching Department of Computer Science Sant Gadge Baba Amravati University, Amravati.
He has completed BE, ME and PhD in computer Science & Engg in 1990, 1997 and 2009 respectively. He has around 35 publications to his credit in International/National Journal and Conferences. His area of interest includes Soft Computing and Real-time system.

Dr. V. M. Thakare
Dr. Vilas M. Thakare is Professor and Head in Post Graduate department of Computer Science and engg, Faculty of Engineering & Technology, SGB Amravati university, Amravati. He is also working as a co-ordinator on UGC sponsored scheme of e-learning and m-learning specially designed for teaching and research. He is Ph.D. in Computer Science/Engg and completed M.E. in year 1989 and graduated in 1984-85.