Expert Systems With Applications 103 (2018) 92–105

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Senti-N-Gram: An *n*-gram lexicon for sentiment analysis

Atanu Dey*, Mamata Jenamani, Jitesh J. Thakkar

Department of Industrial and Systems Engineering, Indian Institute of Technology, Kharagpur, Kharagpur 721302, India

ARTICLE INFO

Article history:
Received 4 August 2017
Revised 28 February 2018
Accepted 2 March 2018
Available online 6 March 2018

Keywords:
Senti-N-Gram
n-grams
Lexicon
Sentiment analysis
Automatic sentiment score calculation
Automatic *n*-gram construction

ABSTRACT

Sentiment analysis helps evaluating the performance of products or services from user generated contents. Lexicon based sentiment analysis approaches are preferred over learning based ones when training data is not adequate. Existing lexicons contain only unigrams along with their sentiment scores. It is observed that sentiment *n-grams* formed by combining unigrams with intensifiers or negations show improved results. Such sentiment *n-gram* lexicons are not publicly available. This paper presents a methodology to create such a lexicon called *Senti-N-Gram*. Proposed rule-based approach extracts the n-grams sentiment scores from a random corpus containing product reviews and corresponding numeric rating in five-point scale. The scores from this automated procedure are compared with that of the human annotators using *t-test* and found to be statistically equivalent. The paper also proposes a sentiment classification methodology by using a ratio based approach based on counts of positive and negative sentences of a document. When used *Senti-N-Gram* lexicon, proposed method outperforms well-known unigram-lexicon based approach using *VADER* and an *n-gram* sentiment analysis approach *SO-CAL*.

1. Introduction

Sentiment analysis deals with automatic extraction of people's opinions or emotions towards products and services from user generated content (Liu, 2011; Liu & Chen, 2015; Pang, Lee et al., 2008). It has become particularly important with the introduction of Web 2.0 which enables the users to express their views on objects through consumer forums, social media and e-governance portals (Speriosu, Sudan, Upadhyay, & Baldridge, 2011; Tang, Wei, Qin, Zhou, & Liu, 2014). Both companies and government are using these user generated contents to evaluate the performance of products or services (Arunachalam & Sarkar, 2013; Mittal, Goel, & Jain, 2016). They use sentiment analysis tools for identifying the polarity of the contents (Liu & Chen, 2015; Peng, Zuo, & He, 2008). Such tool is useful for making purchase decision by user and product improvement by manufacturer (Bag, Tiwari, & Chan, 2017; Peng et al., 2008).

There are two broad approaches for calculating the sentiment of a text document: *rule-based* and *machine learning based*. The machine learning based approaches classify the user-generated contents into positive or negative classes using some commonly used classifiers such as Naïve Bayes (NB) (McCallum, Nigam et al., 1998), Maximum Entropy (ME) (Nigam, Lafferty, & McCallum, 1999), Support Vector Machine (SVM) (Hsu, Chang, Lin et al., 2003) etc. The classifiers need sizable labeled datasets for training and testing (Dey, Jenamani, & Thakkar, 2017; Liu & Chen, 2015; Pang, Lee, & Vaithyanathan, 2002; Tripathy, Agrawal, & Rath, 2016). Generating such gold-standard labeled datasets requires human annotators and which is expensive in terms of money and time. Therefore, though these approaches give better result for classification problem, the rule-based approaches are often preferred where training datasets are hard to obtain (Baccianella, Esuli, & Sebastiani, 2010; Hutto & Gilbert, 2014; Turney, 2002; Hogenboom, Heerschop, Frasincar, Kaymak, & de Jong, 2014).

The rule-based approaches evaluate the sentiments using publicly available lexicons (Baccianella et al., 2010; Cambria, Havasi, & Hussain, 2012; Hutto & Gilbert, 2014; Stone, Dunphy, & Smith, 1966). Table 1 shows the list of most frequently used such lexicons in a chronological manner. New lexicons get developed to resolve the issues such as lack of frequent uses words of that time, slangs and improvement of the sentiment scores. While the first four of the list are used for varieties of Natural Language Processing Applications, the next three are developed specifically for sentiment analysis task. So far, VADER is the newest and higher performing lexicon among all for sentiment lexicons (Hutto & Gilbert, 2014). It may be noted that, all these lexicons are for unigrams.

There have been some efforts for using n-gram in lexicon based sentiment analysis (Moreo, Romero, Castro, & Zurita, 2012; Satthar, 2015; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Hogenboom, Van Iterson, Heerschop, Frasincar, & Kaymak, 2011; Jia, Yu, & Meng, 2009). However, such approaches do not propose to create *n*-

* Corresponding author.
E-mail addresses: atanu.dey@iitkgp.ac.in (A. Dey), mj@iem.iitkgp.ernet.in (M. Jenamani), jt@iem.iitkgp.ernet.in (J.J. Thakkar).

https://doi.org/10.1016/j.eswa.2018.03.004
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

Table 1
Comparison among different unigram-lexicons for sentiment analysis.

Lexicon	Procedure	Contributors	Remarks
General Inquirer (GI)	Manual	Stone et al. (1966)	• This lexicon contains more than 11K words classified into one or more of 183 categories. • Sentiment analysis researchers generally focus on 1915 positive labeled words and 2291 negative labeled words.
WordNet	Manual	Fellbaum (1998)	• WordNet is a database of English words that are linked together by their semantic relationships. • It is like a supercharged dictionary/thesaurus with a graph structure.
ANEW	Manual	Bradley and Lang (1999)	• Affective Norms for English Words (ANEW) lexicon provides a set of normative emotional ratings for 1034 words. • Unlike LIWC or GI, the words in ANEW have been ranked in terms of their pleasure, arousal and dominance.
LIWC	Manual	Pennebaker, Francis, and Booth (2001)	• Linguistic Inquiry and Word Count (LIWC) is annotated and both internally and externally validated in a process spanning more than one decade of work by psychologists, sociologists, and linguists.
SentiWord- Net	Semi-Automatic	Baccianella et al. (2010)	• SentiWordNet is an extension of WordNet in which 147,306 synsets are annotated with three numerical scores relating to positivity, negativity and neutrality.
SenticNet	Semi-Automatic	Cambria et al. (2012)	• This is a publicly available semantic and affective resource for concept level opinion mining. • SenticNet is constructed by means of sentic computing. It contains 14,244 common sense concepts.
VADER	Manual	Hutto and Gilbert (2014)	• VADER contains 7500 semantic words and each word has either positive or negative semantic intensity within the range of -4 to $+4$. • This dictionary does not contain any neutral word. Performance of this dictionary is considerable compared to other existing unigram dictionaries.

gram dictionaries. As a result, every time an *n*-gram is used, its score needs to be recalculated; which in turn increases the computational time. Therefore, though *n*-gram approach shows improved performance compared to unigram approaches (Hutto & Gilbert, 2014), not much work has been done in this area. Our work extends this field by creating an n-gram dictionary.

Creating a lexicon can be manual, semi-manual or automatic. The dictionaries such as GI, WordNet, ANEW, LIWC, VADER are created using manual method where human subjects are involved. Such manual processes are expensive in terms of both time and cost. As the name indicates semi-manual methods combines both human annotators and algorithms to build the dictionary. There have been some efforts for semi-manual (SentiWordNet Baccianella et al., 2010 and SenticNet Cambria et al., 2012) and automatic (Almatarneh & Gamallo, 2017; Deng, Sinha, & Zhao, 2017; Tan & Wu, 2011) sentiment lexicons creation. However, most of these automatic efforts are for domain specific sentiment analysis task and deal with unigrams except for Tan and Wu (2011) and Almatarneh and Gamallo (2017). Though, these two authors give the scores for few multi-words, the list is limited and constrained by the domain on which it is described. To this end, we propose an automatic procedure for creating a general purpose *n*-gram lexicon.

Precursor of our effort is availability of large datasets which contain both textual reviews and corresponding numeric ratings from consumers. We hypothesize that the numeric rating may be used as a replacement for the judgment by human subjects to extract sentiment score. Exploiting this data source, we propose an approach for creating a sentiment lexicon that automatically calculates the n-gram scores considering a lists of intensifiers, negations and semantic unigrams. Automation is realized with the help a corpus containing more than 1,00,000 customer reviews along with intensity rating in five-point scale following a rule based approach. We call this *n*-gram dictionary as *Senti-N-Gram*. We also propose a method for calculating sentiment score at sentence level. We conduct few experiments with two benchmark datasets to study the effectiveness of the proposed lexicon and the method for sentiment calculation. For comparison purpose, we use VADER – a unigram dictionary (Hutto & Gilbert, 2014), and SO-CAL – a method for finding n-gram sentiment score (Taboada et al., 2011). The comparison of our sentiment classification method with few recent ones shows that our method in combination with the *Senti-N-Gram* lexicon perform the better.

Our distinct contributions for the paper can be summarized as follows:

- To the best of our knowledge, publicly available domain independent n-gram lexicons do not contain the semantic scores (Lin et al., 2012). We are probably the first to create such a dictionary.
- We propose to use a random corpus with text reviews and numeric ratings as a replacement of human annotators for automatic dictionary creation. The results when compared with the manual annotators show statistical equivalence of the scores from both the sources.
- Existing *n*-gram based sentiment analysis methods depend on the publicly available unigram lexicons and maintain a list of intensifier with some pre-specified scores. When a sentiment *n*-gram is encountered in the text the score is calculated on-line. Our method proposes an algorithm to extract the senti-*n*-grams used in a publicly available corpus and relate the reviews with numeric rating to extract the score. To the best of our knowledge, this probably the first fully automatic score calculation algorithm to create a domain independent *n*-gram sentiment dictionary. We statistically prove that the proposed algorithm can indeed replace human experts used for annotation.
- Our proposed method for sentiment classification of consumer reviews is unique from the existing methods. The method uses the ratio of total positive and negative sentences as a metric for overall document level evaluation; whereas, other existing methods add up the sentence level score and consider their mean as the document level sentiment. Experiments show our method outperforms others when used with the proposed *Senti-N-Gram* dictionary. In particular, we compare with an *n*-gram sentiment analysis approach SO-CAL and a unigram lexicon based approach using VADER.

The rest of the paper is structured as follows: Section 2 describes the related work and the background of this work. Section 3 presents the proposed framework for creating *Senti-N-Gram* lexicon. Section 4 shows the proposed algorithm for score calculation and refinement of *Senti-N-Grams*. Section 5 demonstrate the score calculation of *Senti-N-Grams*. Section 6 describes