International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

Using Hybrid Approach Analyzing Sentence Pattern by POS Sequence over Twitter

Avinash B. Surnar[1], Sachin N. Deshmukh[2]

[1,2] Department of Computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India

---

Abstract - *Twitter is one of the largest social media providing large amount of public data every day in the form of short messages known as Tweets. Researcher study Twitter for determining public opinion about issues, entity, product reviews, movie reviews or elections which is known as Sentiment Analysis. In recent years many methods as well as techniques are proposed regarding to Sentiment Analysis. This paper proposes POS sequence to determine the sentence pattern or word sequence from tweets in two forms subjectivity and polarity. To determine the POS Sequence, Information Gain is considered for 2-tags sequence and 3-tags sequence. The results shows distinguishable sentence pattern groups of positive and negative tweets. POS sequence approach used in this paper can improve accuracy.*

Key Words: Sentiment Analysis, Polarity, Subjectivity, POS Sequence, Lexicon Analysis.

1. INTRODUCTION

The emergence of social media combined with micro blogging services easy-to-use features have dramatically changed people's life with more and more people sharing their thoughts, expressing opinions. Twitter is social networking service where users post and interact with messages. Twitter is popular for its massive spreading of instant messages (i.e. tweets) and the nature of freedom. Twitter may provision for an excellent channel for opinion creation and presentation. So these tweets are used for analyzing opinions about current issues, product reviews, movie reviews, elections. Sentiment mining in twitter can be used for real time applications like marketing [1].

The process of identifying and categorizing opinions expressed in a piece of text is known as Sentiment Analysis. Sentiment analysis is also called as opinion mining which is the field of study as well as analyzing people's opinions, sentiments, appraisals, evaluations, attitudes, and emotions regarding entities like products, services, organizations, individuals, issues, topics, events for knowledge discovery. This field refers to a broad area of natural language processing, computational linguistic, and text mining. Typically, the goal is to determine the polarity of natural language texts. The area of Sentiment Analysis intends to comprehend the opinions and distribute them into the categories like positive, negative. Sentiment analysis can be performed by subjectivity classification and polarity classification. In the world Textual information can be classified into two main categories as facts and opinions. Facts are also known as objective statements about entities and events. Opinions are also known as subjective statements that reveal people's sentiments or views about the entities and events.

Part of Speech is linguistic category of words which is generally defined by the syntactic. We use combination of POS tags in order to investigate the pattern of word combination that commonly appears in tweet containing sentiment. This experiment is performed by using sequence of 2 and 3 tags. Information Gain is calculated and then uses the *top-k* sequences [2]. In addition we also perform supervised classification in which we incorporate POS sequence with previous method in Twitter sentiment analysis.

2. RELATED WORK

The first investigation of tweet sentiment was done by Go et al. in which they utilized emoticons to annotate tweet with sentiment label [3]. The next study by Agarwal et al. used manually annotated tweets with sentiment and perform unigram model to do classification [4]. Bandersky et al. used POS Sequence as one of features in detecting memorable quote from structured document like book. Specifically, they used Information Gain to select top-*i* sequence and then perform a supervised quotable phrase detection using other lexical and syntactic features [5]. Mukherjee et al. has also proposed POS sequence as feature in gender classification of blog authors. The main idea of their algorithm is to perform a level-wise search for such patterns, which are POS sequences with minsup and minadherence [6].

Pak and Paroubek used a dataset formed of collected messages from Twitter. This paper shows how to automatically collect a corpus for sentiment analysis and opinion mining purposes. This work is able to determine positive, negative and neutral sentiments of documents. The classifier is based on the multinomial Naïve Bayes classifier that uses N- gram and POS-tags as features [7]. Xia et al. used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets

as Part-of-speech information and Word-relations as well as three base classifiers i.e. Naive Bayes, Maximum Entropy and Support Vector Machines [8].

3. POS SEQUENCE AS FEATURE

POS sequence is defined as a series of several tags which are limited to certain number of tags. For example, a sentence "I went to temple yesterday", its POS sequence is PRP-VBD-TO-NN-ADV. It can produce 4 sequences of 2-tags: PRP-VBD, VBD-TO, TO-NN and NN-ADV as well as 3 sequences of 3-tags: PRP-VBD-TO, VBD-TO-NN, TO-NN-ADV. Bandersky et al used Information Gain to select the top-i sequences.

$$IG(X,Y) = H(X) - H(X|Y) \quad (1)$$

$$H(X) = -P(x)\log_2 P(x) \quad (2)$$

Where X indicates the presence or absence of POS sequence in current tweet, and Y indicates the type of tweet. It can be positive or negative for polarity classification and subjective or objective for subjectivity classification.

The preprocessing contains removing URL and removing non-alphabetic symbols, removing retweets, converting tweet into lowercase character, removing stopwords and removing punctuation. Then POS tagger is applied for determining tags. After that information gain is calculated for all sequences. Finally, we select the top of-*i* sequences for Sentiment Analysis.

4. EXPERIMENT

The experiment is conducted on Standford Twitter Sentiment (STS) which was used by Go et al [9]. Sanders dataset, Health Care Reform (HCR), Obama-McCain Debate (OMD) which were used by Speriosu et al. and International Workshop Sem-Eval 2013 (SemEval) data.

Each tweet in these datasets includes a positive, negative tag. Polarity classification is done by only using positive and negative label, while our subjectivity classification considers neutral tweet as objective and positive/negative tweet as subjective.

NLTK is used as POS Tagger to build data in POS tag form. Then tweets are extracted to several sequences of *n*-tags based on steps shown in figure i.e. 2- tags and 3-tags. After that the top-100 of sequences of n-tags are selected based on Information Gain. Finally SVM is applied to the extracted data. This can be performed on all datasets used. Then select top 10 sequences for analysis of sentence pattern.

Table -1: Dataset Statistics

Dataset	No. Of Tweets
STS	2628
Sanders	4800
HCR	620
OMD	1994
SemEval	1620

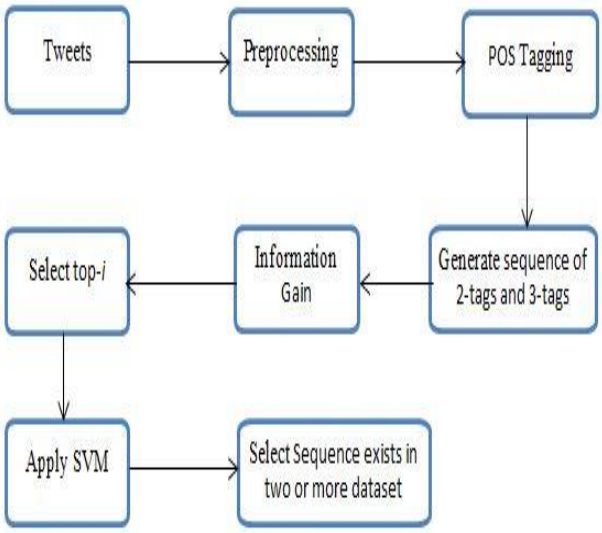

Fig -1: Sequence Extraction with SVM

In addition, sentiment classification is performed using top-100 POS sequences determined by Information Gain. As baseline, AFINN [10], a lexicon containing 5012 English words and constructed based on the *Affective Norms for English Words* lexicon (ANEW) proposed by Bradley and Lang [11] as well as MPQA [12], lexicons contain 13770 English words.

After applying SVM weighting to all datasets in forms of sequence variations of top-10 POS sequences are produced. To perform analysis, a sequence which exists in two or more datasets is selected. The result is shown in Table. It also provides the frequency of POS sequence of word combination between two classes. In these tables, result of sequence containing 2 and 3 tags is given.

To determine the performance of POS sequence in sentiment classification, we compared AFINN lexicon with the incorporation of POS sequence. AFINN lexicon is used by extracting tweet into two main features called APO (AFINN Positivity) and ANE (AFINN Negativity). The combination of AFINN and POS sequence are able to boost the accuracy of AFINN lexicon. In this, MPQA lexicon is used for the comparison of AFINN and MPQA accuracy.

POS sequences containing 2-tags (JJ-NN, PRP-VBP-JJ, RB-VB, and RB-NN), POS sequences containing 3-tags (NN-DT-NN, NN-NN-IN) shown in table1 and table2 respectively. Unlike sequence of 2-tags and 3-tags, results of 5-tags sequence are difficult to interpret. It is caused by more combination of POS sequences for higher *n* value.

Table -2: POS Sequence of 2-tags with Frequency

Sequence of 2-Tags	Description	Dataset	Frequency	
			Positive	Negative
JJ-NN	Adj-Noun	1	511	466
NN-DT	Noun-Det	1	138	164
RB-VB	Adverb-Verb	2	189	210
NN-PRP	Noun-Pronoun	4	105	156
VBZ-DT	Verb-Det	3	324	249
PRP-VBD	Pronoun-Verb	2	227	179
RB-NN	Adverb-Noun	5	138	148
VBP-JJ	Verb-Adj	4	230	152

Table -3: POS Sequence of 3-tags with Frequency

Sequence of 3-Tags	Description	Dataset	Frequency	
			Positive	Negative
DT-NN-IN	Det-Noun-Conj	1	177	166
NN-NN-IN	Noun-Noun-Conj	2	165	138
VBD-NN-NN	Adverb-Verb	4	146	120
NN-DT-NN	Noun-Det-Noun	4	105	156
NN-NN-VBP	Noun-Noun-Verb	5	94	63
MD-RB-VB	Modal-Adverb-Verb	1	50	89

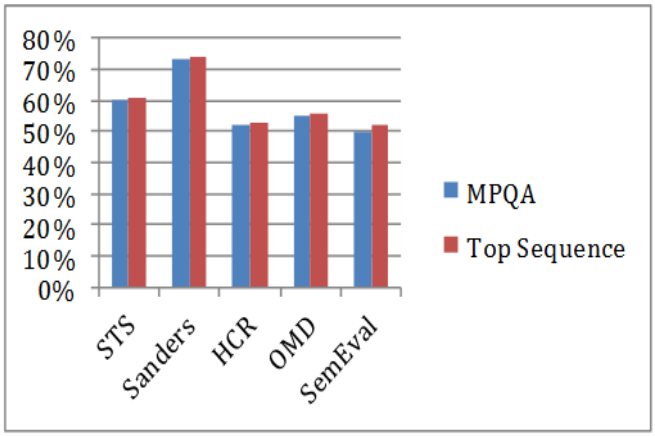

Chart -1: MPQA with Top Sequence of 2-tags

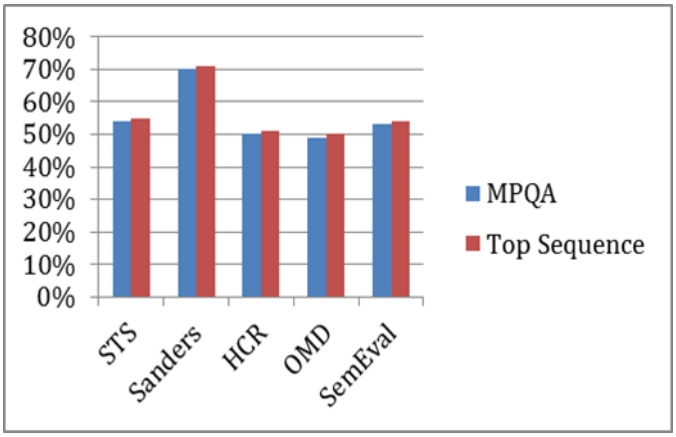

Chart -2: MPQA with Top Sequence of 3-tags

5. CONCLUSIONS

The use of POS Sequence in Sentiment Analysis over Twitter to achieve the top POS sequence in sentence pattern, we performed the study in variations of POS sequence 2-tags and 3-tags. In addition, we performed sentiment classification by including AFINN Lexicon as well as MPQA Lexicon and POS Sequence. In our first experiment, the results reveal that subjective tweets tend to have word combinations consisted of adverb and adjective. The results show that features of POS sequence can improve the accuracy in incorporation between AFINN and POS sequence. As AFFIN and MPQA are incorporated MPQA provides better accuracy. So that POS sequence can be efficiently used for performing Sentiment Analysis over Twitter. Because of the informal nature of tweets, spelling of the words can vary a lot. Because the available lexicons are mainly based on the words from pure English language, there is problem when finding for prior polarity features. Therefore need to create more sophisticated lexicon dictionary according to the words used in the twitter messages.

REFERENCES

[1] Amit Narote, sohail Shaikh, saville periera, nitin jadhav, platini Rodrigues, "Sentimental analysis on Twitter", International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137.

[2] Fajri Koto, and Mirna Adriani, "The Use of POS Sequence for Analyzing Sentence Pattern in Twitter Sentiment Analysis," International Conference on Advanced Information Networking and Applications Workshops 2015.

[3] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision." In CS224N Project Report, Stanford, pp. 1-12, 2009.

[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data". In Proc. of the Workshop on Languages in Social Media, pp. 30-38, 2011.

[5] M. Bendersky, and D. A. Smith, "A dictionary of wisdom and wit: Learning to extract quotable phrases." In Proc. NAACL-HLT 2012, pp. 69, 2012.

[6] A. Mukherjee, and B. Liu, "Improving Gender Classification of Blog Authors". In Proceeding of the Conference on Empirical Methods in Natural Language Processing, pp. 207-217, 2010.

[7] A.Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", the Seventh Conference on International Language Resources and Evaluation, pp.1320–1326, 2010.

[8] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[9] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," In CS224N Project Report, Stanford, pp. 1-12, 2009.

[10] F. A. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", Available at http://arxiv.org/abs/1103.2903, 2001.

[11] M. M. Bradley, and P. J. Lang, "Affective norms for English words (ANEW): Instruction ma-nual and affective ratings", Technical Report C-1, the Center for Research in Psychophysiology, University of Florida, 1999.

[12] Multi Perspective Question Answering (MPQA). Online Lexicon "http://www.cs.pitt.edu/mpqa/subj_lexicon.html".