World Applied Sciences Journal 29 (Computer Sciences, Engineering and Its Applications): 131-135, 2014
ISSN 1818-4952
© IDOSI Publications, 2014
DOI: 10.5829/idosi.wasj.2014.29.csea.2242

Supervised Approach to Extract Sentiments from Unstructured Text

K.P. Thooyamani, R. Udayakumar and V. Khanaa

School of Computing Science, Bharath University, Chennai-73, India

Abstract: Sentiment Analysis is a two level task. The first one is Identifying Topic and the second is, classifying sentimentrelated to that topic. Sentiment Analysis starts with “What other people thinks?”. Sentiment Extraction deals with the retrieval of the opinion or mood conveyed in a block of Unstructured text in relation to the domain of the document being analyzed. Although a lot of research has gone in the NLP, machine learning and web mining community on extracting structured data from unstructured sources, most of the4 proposed methods depend on tediously labeled unstructured data. The World Wide Web has been dominated by unstructured content and searching the web has been based on techniques from Information Retrieval. Supervised learning algorithm analyzes the training data and produces an inferred function which is called classification.

Key words: Sentiment Extraction · Rating words · Polarity detection · Customer feedback

INTRODUCTION

The World Wide Web is growing at an alarming rate not only in size but also in the types of services and contents provided. Individual users are participating more actively and are generating vast amount of new data.These new web contents include customer reviews and blogs that express opinions on products and services which are collectively referred to as customer feedback data on the web. As customer feedback on the web influences other customer’s decisions, these feedbacks have become an important source of information for businesses to take into account when developing marketing and product development plans [1, 2].

Sentiment Extraction is a relatively growing field of research fuelled by the growing ubiquity of the Internet coupled with the huge volume of data being generated in it in the form of review sites, web logs and wikis [3]. It so happens that over eighty percent of data on the Internet is unstructured and is available from feedback fields in survey, blogs, wikis and so on. This huge volume of data might posses potential profitable business related information, which when extracted intelligently and represented sensibly, can be a mine of gold for a management's R&D, trying to improvise a product based on popular public opinion [4].

Opinion mining refers to a broad area of Natural Language Processing and Text Mining. Most existing approaches apply supervised learning techniques, including Support Vector Machines, Naive Bayes, AdaBoost and others [5]. On the other hand, unsupervised approaches are based on external resources such as WordNet Affect or SentiWordNet [6].

MATERIALS AND METHODS

There are two main techniques for sentiment classification: symbollic techniques and machine learning techniques. The symbollic approach uses manually crafted rules and lexicons,where the machine learning approach uses unsupervised, weakly supervised or fully supervised learning to construct a model from a large training corpus. We proposed a system which uses machine learning techniques instead of symbollic techniques to provide the polarity for sentences present in the world wide web [10].

Machine Learning Techniques

Supervised Methods: In order to train a classifier for sentiment recognition in text classic supervised learning techniques (e.g Support Vector Machines, naïve Bayes Multinomial, Hidden Markov Model)can be used.

Corresponding Author: R. Udayakumar, School of Computing Science, Bharath University, Chennai-73, India.

A supervised approach entails the use of a labelled training corpus to learn classification function. The method that in the literature often yields the highest accuracy regards a Support Vector Machine classifier. They are the ones we used in our experiments described below [8].

Support Vector Machines(SVM): SVM operate by constructing a hyperplane with maximal Euclidean distance to the closest training examples. This can be seen as the distance between the separating hyperplane and two parallel hyperplanes at each side, representing the boundary of the examples of one class in the feature space. It is assumed that the best generalization of the classifier is obtained when this distance is maximal. If the data is not separable, a hyperplane will be chosen that splits the data with the least error possible.

Naive Bayes Multinomial (NBM): A naive Bayes classifier uses Bayes rule (which states how to update or revise believes in the light of new evidence) as its main equation, under the naive assumption of conditional independence: each individual feature is assumed to be an indication of the assigned class, independent of each other [9]. A multinomial naïve Bayes classifier constructs a model by fitting a distribution of the number of occurrences of each feature for all the documents.

Hidden Markov Model (HMM): We present a novel probabilistic method for topic segmentation on unstructured text. One previous approach to this problem utilizes the hidden Markov model (HMM) method for probabilistically modeling sequence data [7]. The HMM treats a document as mutually independent sets of words generated by a latent topic variable in a time series. We extend this idea by embedding Hofmann's aspect model for text [5] into the segmenting HMM to form an aspect HMM (AHMM). In doing so, we provide an intuitive topical dependency between words and a cohesive segmentation model. We apply this method to segment unbroken streams of New York Times articles as well as noisy transcripts of radio programs on Speech about, an online audio archive indexed by an automatic speech recognition engine [11, 12].

Challenges: Most of the challenges pertaining to SE arise from the vagaries of natural language. Some critical challenges that people face in this do -main are elucidated below.

- Most of the approaches depend on a rating word in determining sentiment of a phrase. But cases exit where phrases express contextual sentiments without a rating word being used. For example, consider the sentence “Steve Waugh is not a cricketer but can be a peanut seller”.The sentence conveys a strong negative sentiment but no rating words have been used.
- Sarcasm might be intended but might not be interpreted, leading to terribly wrong results. For example, consider the phrases “Terrorists are really nice guys.They rid the innocent of their pains and send them to the lotus feet of god”. The example shows a phrase that will anchor terrorists with a positive polarity, a complete irony!
- Synonym databases and lexicons are never exhaustive and tend to give out of context results, a direct consequence of the underlying complexity involved in a natural language.
- Double negations can lead to unexpected results that are seldom accounted for. As an example, the statement “It ain’t no good” conveys a negative sentiment inspite of the double negation.
- Anaphora resolution, i.e., attaching pronouns to nouns is an important challenge in the SE domain.
- The most important problem is that the process of sentiment extraction is not generic but highly domain specific. The lexicons and other linguistic resources used should be domain relevant in order to get meaningful results. In addition, these should constantly be tweaked (probably with machine learning techniques) to be in tune with newer developments in the concerned domain.
- There exists the problem of subectivity and neutral texts [13]. One must have detectors to remove portions of texts which do not convey any sentiments to improve accuracy of the engine.
- A major problem lies in quantifying the polarities of the rating words, intensifiers, nagators and the computed sentiment. The scale of polarity adapted and the mathematical results that follow from computations have to be mapped to something significant and tangible to the end user.
- A significant factor to be noted is that entities are generally recognized from statistical machine learning algorithms which just give out probabilistic results. Therefore there are good chances of a phrase being tagged with a wrong or an out of context entity.

The Proposed System: It consists of two basic components: word sense disambiguation and determination of polarity. The first, given an opinion, determines the correct senses of its terms and the second, for each word sense determines its polarity and from them gets the polarity of the opinion [14].

Firstly, a preprocessing of the text is carried out including sentence recognizing, stop word removing, part-of-speech tagging and word stemming by using the Tree Tagger tool.

Word Sense Disambiguation (WSD) consists on selecting the appropriate meaning of a word given the context in which it occurs. For the disambiguation of the words, we use the method proposed in, which relies on clustering as a way of identifying semantically related word senses.

In this WSD method, the senses are represented as signatures built from the repository of concepts of WordNet. The disambiguation process starts from a clustering distribution of all possible senses of the ambiguous words by applying the Extended Star clustering algorithm [15]. Such a clustering tries to identify cohesive groups of word senses, which are assumed to represent different meanings for the set of words. Then, cluster that match the best with the context are selected. If the selected clusters disambiguate all words, the process stops and the senses belonging to the selected clusters are interpreted as the disambiguateing ones.

Otherwise, the clustering are performed again (regarding the remaining senses) until a complete disambiguation is achieved.Once the correct sense for each word on the opinion is obtained, the method determines its polarity regarding the sentiment values for this sense in *SentiWordNet* and the membership of the word to the *Positiv* and *Negativ* categories in GI. It is important to mention that the polarity of a word is forced into the opposite class if it is preceded by a valence shifter (obtained from the *Negate* category in GI).

Naive Bayes Classification Model: The classification process is done by Naive Bayes Classification algorithm. It assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature given the class variable. For some type of probability models, nBayes classifiers can be trained very efficiently in a Supervised learning setting. A supervised approach entails the use of a labeled training corpus to learn a certain classification function.

Bayes Theorem for Plain English:

$$\text{Posterior} = \frac{\text{Prior * Likelihood}}{\text{Evidence}}$$

- Posterior-Probability of the observed text,
- Prior-The initial probability before seeing any evidence,
- Likelihood-Probability of observing sample,
- Evidence-Class label is unknown.

Finally, the polarity of the opinion is determined from the scores of positive and negative words it contains. To sum up, for each word *w* and its correct sense *s*, the positive (*P*(*w*)) and negative

(*N*(*w*)) scores are calculated as:

- P(w) ={ *otherwise*
- *category in GI*
- *if w belongs to the Positiv*
- *positive value of s in SentiWN*
- *positive value of s in SentiWN*
- *P w* ……………………………. (1)

N(w)=*otherwise*
category in GI

- *if w belongs to the Negative*
- *negative value of s in SentiWN*
- *negative value of s in SentiWN*
- *N w* …………………………… (2)

Finally, the global positive and negative scores (*Sp*, *Sn*) are calculated as:

Sp= ∑ p(w)   Sn =∑N(w)

W :p(w) > N(w) w:N(w)>p(w) … (3)

If *Sp* is greater than *Sn* then the opinion is considered as positive. On the contrary, if *Sp* is less than *Sn* the opinion is negative. Finally, if *Sp* is equal to *Sn* the opinion is considered as neutral.

Opinion Summarization: Unlike traditional text summarization that tries to construct short text which efficiently expresses the subject of the original long text, opinion sum-marization aims to give the overall sentiment of a large amount of reviews or other form of opinion resources at various granularities. It is relatively trivial that sentiment classification may be one subtask of opinion summarization. For instance, generally each review is classified and then the ratio of the positives and negatives is suggested as the overall favorableness on the product.

Nevertheless we concentrate on how the overall sentiment of each feature of a product is summarized. We do this by looking into several opinion mining systems. In the system we have examined, product features are extracted and then sentiment of each feature is assigned.

Then these are summarized and presented in various forms. Most of the current systems extract product features largely based on the statistical approach. On the contrary, various methods are used for assigning sentiment to the extracted features: PMI method, supervised classification method and syntactic analysis. Some of the OM systems use linguistic resources which contain sentiment lexicons and others use star ratings or thumbs up/down icons instead.

Overall Architecture:

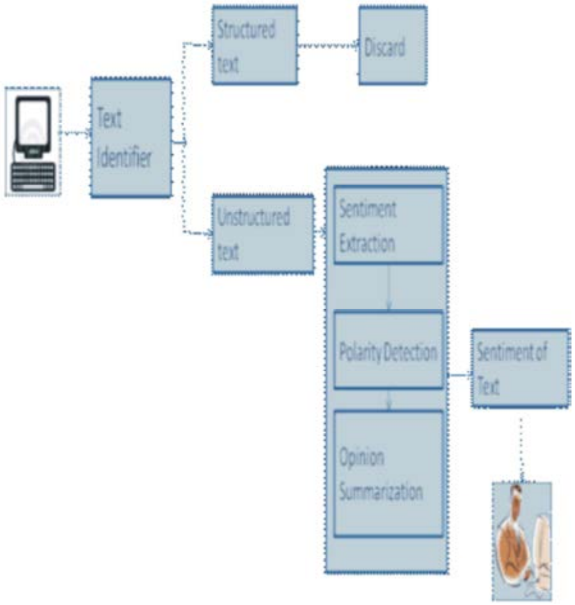

Fig 1**:**

Tools Used:

- *Word Sense Disambiguation(WSD)*
- *Word Net*
- *SentiWordNet*
- *General Inquirer*

Word Sense Disambiguation(WSD): It consists on selecting the appropriate meaning of a word given the context in which it occurs. For the disambiguation of the words, we use the method proposed in (Anaya-Sánchez *et al*., 2006), which relies on clustering as a way of identifying semantically related word senses.

Word Net: WordNet, adjectives are organized into bipolar clusters and share the same orientation of their synonyms and opposite orientation of their antonyms. To assign orientation of an adjective, the synset of the given adjective and the antonym set are searched.

If a synonym/antonym has known orientation, then the orientation of the given adjective could be set correspondingly. As the synset of an adjective always contains a sense that links it to the head synset, the search range is rather large. Given enough seed adjectives with known orientations, the orientations of all the adjective words can be predicted.

Senti Word Net: SentiWordNet (Esuli and Sebastiani, 2006) is a lexical resource for opinion mining. Each synset in WordNet has assigned three values of sentiment: positive, negative and objective, whose sum is 1. It was semi-automatically built so all the results were not manually validated and some resulting classifications can appear incorrect.

General Inquirer: General Inquirer (GI) (Stone *et al*., 1966) is an English dictionary that contains information about the words. For the proposed method we use the words labelled as positives, negatives and negations (Positiv, Negativ and Negate categories in GI).

From the Positiv and Negativ categories, we build a list of positive and negative words respectively. From the Negate category we obtain a list of polarity shifters terms (also known as valence shifters).

CONCLUSION

In this paper, a new method for Sentiment Extraction of Unstructured Text was introduced to determine the polarity and summarize the text efficiently. Its most important novelty is the use of WordNet and Word Sense Disambiguation tools together with standard external resources for determining the polarity of the opinions. These resources allow the method to be extended to other languages and be independent of the knowledge domain.

Samples:

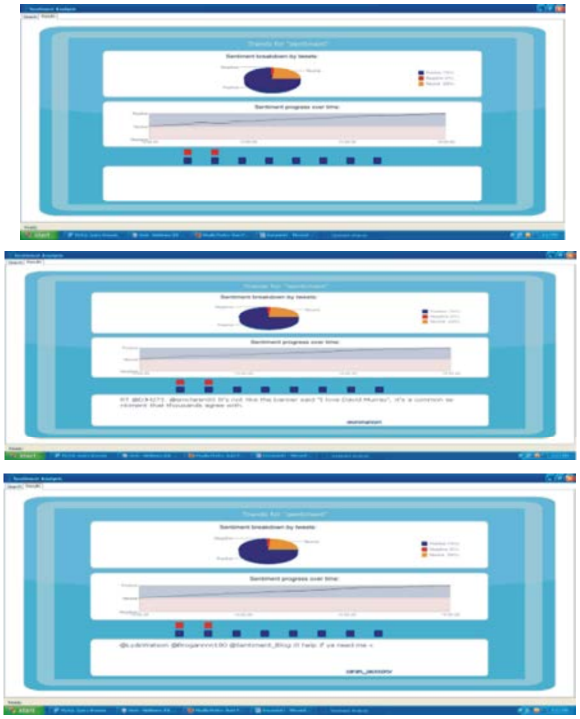

Fig 2:

REFERENCES

1. Automatic Sentiment Analysis in On-line Text Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens Katholieke Universiteit Leuven, Tiensestraat 41 B-3000 Leuven, Belgium.
2. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews Peter D.Turney Institute for Information Technology National Research Council of Canada Ottawa, Ontario, Canada, K1A 0R6peter.turney@nrc.ca
3. OPINION POLARITY DETECTION, Using Word Sense Disambiguation to Determine the Polarity of Opinions Tamara Martín-Wanton, Aurora Pons-Porrata Center for Pattern Recog-nition and Data Mining, Universidad de Oriente, Patricio Lumumba s/n, Santiago de Cuba, Cuba tamara@cerpamid.co.cu, aurora@cerpamid.co.cu
4. Osgood, C.E., G.J. Suci and P.H. Tannenbaum, 1971. The Measurement of Meaning. University of Illinois Press, [1957].
5. Biber, D. and E. Finegan, 1989. Styles of stance in english: Lexical and grammatical marketing of evidentiality and affect. Text, 9: 93-124.
6. Wallace, A.F.C. and M.T. Carson, 1973. Sha-ring and diversity in emotion terminology. Ethos., 1(1): 1-29.
7. Hatzivassiloglou, V. and J. Wiebe, 2000. Effects of adjective orientation and gradability on sentence subjectivity, Proceedings of the 18th International Conference on Computational Linguistics, ACL, New Brunswick, NJ.
8. Fellbaum, C. (ed.), 1998. Wordnet: An electronic lexical database, Language, Speech and Communication Series, MIT Press, Cambridge.
9. Kamps, J., M. Marx and R.J. Mokken and M. De Rijke, 2004. Using WordNet to measure semantic orientation of adjectives. LREC, volume IV, pp: 1115-1118.
10. Mulder, M., A. Nijholt, Den M. Uyl and P. Terpstra, 2004. A lexical grammatica implementation of affect, Proceedings of TSD-04, the 7th International Conference Text, Speech and Dialogue, Lecture Notes in Computer Sci-ence, vol. 3206, Springer-Verlag, Brno, CZ, pp: 171-178.
11. Dave, K., S. Lawrence and D.M. Pennock, 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.In Proceedings of WWW-03, 12th International Co-nference on the World Wide Web, ACM Press, Budapest, HU, pp: 519-528.
12. Pedersen, T., 2001. A decision tree of bigrams is an accurate predictor of word sense. In Proceed-ings of the Second Annual Meeting of the North American Chapter of the Association for Comp-utational Linguistics, pp: 79-86.
13. Thooyamani, K.P., V. Khanaa and R. Udayakumar, 2013. A frame work for modelling task coordination in Multi-agent system, Middle-East Journal of Scientific Research, ISSN: 1990-9233, 15(12): 1851-1856.
14. Udayakumar, R., V. Khanna, T. Saravanan and G. Saritha, 2013. Retinal Image Analysis Using Curvelet Transform and Multistructure Elements Morphology by Reconstruction, Middle-East Journal of Scientific Research, ISSN: 1990-9233, 16(12): 1798-1800.
15. Saravanan, T. and R. Udayakumar, 2013. Comparision of Different Digital Image watermarking techniques, Middle-East Journal of Scientific Research, ISSN:1990-9233, 15(12): 1684-1690.