INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN : 2454-9924

Unstructured Data Analysis In Social NetworkUsing BigData

*A.Balu M.E#1. and M.Sreemaa M.E*2*

*#1 PG SCHOLAR, *2 ASSISTANT PROFESSOR*

SURYA GROUP OF INSTITUTIONS

*VIKIRAVANDI,VILLUPURAM***Abstract**—Polarity classification of words is important for applications such as Opinion Mining and Sentiment Analysis. A number of sentiment word/sense dictionaries have been manually or automatically constructed. These sentiment dictionaries have numerous inaccuracies. Besides obvious instances, where the same word appears with different polarities in different dictionaries, the dictionaries exhibit complex cases of polarity inconsistency, which cannot be detected by mere manual inspection. We introduce the concept of polarity consistency of words/senses in sentiment dictionaries in this paper. Sentiment based analysis is the major key in categorizing the user's Feedback. We are using FSM & EEM Algorithm for the Word processing process.The feedback analysis using SVM to improve customer 's experience and brand loyalty by gathering and analyzing customer's feedback. In this not getting feedback using graphical mode,introduce a SVM method so it will give feedback in text and then machine will understand the text and rate for the feedback and bring forum to first rank.

Index Terms—Sentiment analysis, FSM & EEM Algorithm, SVM

1. INTRODUCTION

The opinions expressed in various web and media out-lets (e.g., blogs, newspapers) are an important yard-stick for the success of a product or a government policy. For instance, a product with consistently good reviews is likely to sell well. The general approach of determining the overall orientation (i.e., positive or negative) of a sen-tence/document is by analysis of the orientations of the individual words. Sentiment dictionaries are utilized to facilitate the summarization. Thereare numerous works that, given a sentiment lexicon, analyze the structure of a sentence/document to infer its orienta-tion, the holder of an opinion, the sentiment of the opin-ion, etc.Several domain independent sentiment dictionaries have been manually or (semi)-auto-matically created, We concentrate on the concept of (in)consistency in this paper. We define consistency among the polarities of words/synsets within and across sentiment dictionaries and give methods to check them. sense conveys a positive polarity. Hence, tantalize con-veys a positive sentiment when used with this sense.This solution has an important shortcoming: it generates boolean formulas that have exponential lengths when converting PCC into SAT. We experimentally show that this solution cannot handle words such as give and make which have large numbers of synsets—we left the implementation of this solution running on a quad-core com-puter with 12 GB of memory for a week without ever termi-nating. In this paper, we present a new solution that is proven to generate boolean formulas of polynomial lengths. The new solution can handle all the words in WordNet and it takes only 24 minutes to complete its computations.

2. PROBLEM DEFINITI

As argued above, the polarities of the words in a sentiment dictionary may not necessarily be consistent (or correct). In this paper, we focus on the detection of polarity assignment inconsistency for the words and synsets within and across the

sentiment dictionaries (e.g., OF versus. GI). We attempt to pinpoint the words with polarity inconsistencies. We contend that our approach is applicable to domain dependent sentiment dictionaries, too. We can employ WordNet Domains. WordNet Domains augments WordNet with domain labels such as art, sport, reli-gion and history. Hence, we can project the words and synsets in WordNet according to a domain label and then apply our methodology to the projection.

3. INCONSISTENCY CLASSIFICATION

In this section, we attempt to give a thorough classification with examples of the possible types of polarity inconsisten-cies occurring within and across sentiment dictionaries. Polarity inconsistencies are of two types: input and com-plex. We present them in turn.

3.1 Input Dictionaries Polarity Inconsistency

Input polarity inconsistencies are of two types: intra-dictio-nary and inter-dictionary inconsistencies. The latter are obtained by comparing (1) two SWDs, (2) an SWD with an SSD and (3) two SSDs.

3.1.1 Intra-Dictionary Inconsistency

An SWD to determine the polarity of word

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN : 2454-9924

3.1.2 Intra-Dictionary Inconsistency

An SWD to determine the polarity of word w with part of speech pos. The verb brag has negative polarity according to Definition 2. Such cases simply say that the team who constructs the dictionary believes brag has multiple.

3.2 Complex Polarity Inconsistency

This kind of inconsistency is more subtle and cannot be detected by direct comparison of words/synsets. It consists of a set of words and/or synsets whose polarities cannot concomitantly be satisfied. Recall the example of confute and disprove in OF given. Recall our argu-ment that by assuming that WordNet is correct, it is not pos-sible for the two words to have different polarities: the sole synset, which they share, would have two different polari-ties, which is a contradiction.

3.2.1 WordNet versus Sentiment Dictionaries

The adjective bully is an example of a discrepancy between WordNet and a sentiment dictionary. The word has negative polarity in OF and has a single sense in Word-Net. The sense is shared with the word nifty, which has positive polarity in OF and has a unique sense. By applying Definition 2 to nifty we obtain that the sense is positive, which in turn, by Definition, implies that bully is positive. This contradicts the polarity of bully in OF. According to the Webster dictionary, the word has a sense (i.e., resembling or characteristic of a bully) which has a negative polarities as they do not adopt our dominant sense principle. There are 58 such inconsistencies in GI, OF and AL. QW, a sentiment sense dictionary, does not have intra-inconsistencies as it does not have a synset with multiple polarities.

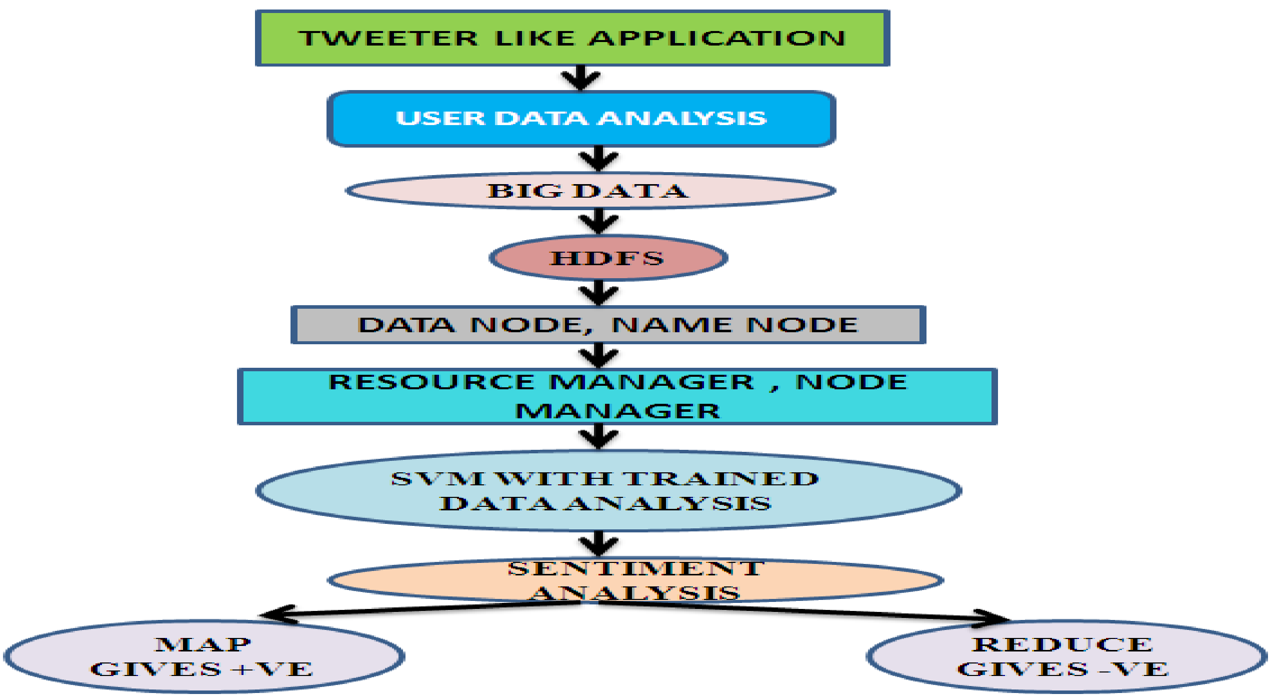
3.2.2 Across Sentiment Dictionaries

We provide examples of inconsistencies across sentiment dictionaries here. Our first example is from SWDs. The adjective comic has negative polarity in AL and the adjec-tive laughable has positive polarity in OF.

4. POLARITY CONSISTENCY CHECKING

To “exhaustively” solve the problem of finding all polarity inconsistencies in a sentiment word dictionary, we propose a solution that reduces an instance of the problem to an instance of CNF-SAT. We can then apply one of the fast SAT solvers to solve our problem. CNF-SAT is a decision problem of determining if there is an assign-ment of True and False to the variables of a Boolean formula F in conjunctive normal form (CNF) such that F evaluates to True. A formula is in CNF if it is a conjunction of one or more clauses, each of which is a disjunction. CNF-SAT is a classic NP-complete problem, but modern SAT solvers are capable of solving many practical instances of the problem.`

5. SYSTEM ARCHITECTURE DIAGRAM



ISSN : 2454-9924

6. COMPLEXITY ANALYSIS OF THE METHODS

In this section, we analyze the complexity of the Boolean formulas generated with the two methods. We start with the analysis of EEM.

6.1 Complexity Analysis of EEM

This method generates a formula, which has double exponential number of clauses in the worst case for a word. The reason is that we first generate a SAT formula that has exponential length in the number of clauses. This formula however is not in CNF and it needs to be converted to CNF. This in general can cause another exponential blow up. Thus, the overall blow up can be dou-ble exponential in the worst case. Because of this, we cannot handle the entire WordNet with it.

6.2 Complexity Analysis of FSM

We now show that the formula generated by FSM is of poly-nomial length in the number of clauses. Suppose that we have a word with m synsets. Corresponding to each internal node in the binary tree, we have k ¼ dlog$_2$ðfreqðwÞÞdþ1 variables representing the binary representation of the number associated with the node. For each such node we have a set of clauses that defines the values of these variables in terms of the values of the variables corresponding to its two children; we also use k additional auxiliary variables that denote the carry bits when the numbers of the children are added. The value of each bit in the sum is defined as a Bool-ean formula of the values of the corresponding bits of the two summands and the carry bit corresponding to the pre-vious bit. Thus, this formula for each bit is a formula over four variables and is obtained directly in CNF. Similarly, we obtain formulas for each carry bit also. The conjunction of all these 2k formulas specifies the values of the bits in the sum in terms of values of bits in its arguments.

6.3 A Hybrid Approach

One drawback of FSM is that it may generate Boolean for-mulas with a large number of variables (thousands). This is particularly the case for words with large number of synsets finish. It required about 7 GB of memory. The hybrid approach has even more efficient, terminating in about 10 minutes. The execution performances of FSM and HYBRID are in steep contrast with that of EEM and we rec-ommend them for use in practice. PicoSAT required the least amount of memory: around 2 GB for both FSM and HYBRID. Its computation time was comparable with that of SAT4j in our experiments.

7. RELATED WORK

There are two lines of work on sentiment polarity lexicon sentiments of adjectives in WordNet by measur-ing the relative distance of a term from exemplars, such as “good or“bad”. The work reports results for adjectives alone. Other approaches use synonyms and antonyms to expand the sets of seeds. Yet another technique is to add all synonyms of a polar word with the same polarity and its antonyms with reverse polarity and, middle, has neutral polarity). QW aims to automatically annotate the synsets (senses) in WordNet. It starts from six synsets with known polarities: “positive”, “negative”, “good”, “bad”, “inferior” and “superior”. These are precisely the synsets that are related to the noun “quality” through the attribute relation in WordNet. It navigates WordNet along the semantic relations defined in WordNet (e.g., hypernym, antonym) and assigns polarities to synsets. If two synsets are assigned conflicting polarities they are discarded. QW does not trace down inconsistencies as we do. Also, they do not assign polarities to words. Finally, the relations in Word-Net do not have well-defined behavior with respect to pre-serving/reversing polarity. Recall the above example of the adjectives advance and middle, which are antonyms, but whose polarities are not reversed.

Unlike SWN, our view is that each synset does not have a degree associated with each polarity. Instead, each synset is 100 percent positive, 100 percent negative or 100 percent neutral.

Machine learning algorithm as well as stochastic algorithms can be employed to classify words into dif-ferent polarities.The differences between our approach and earlier ones, including those that are not WordNet-based to our knowledge, none of the earlier works studied the problem of polarity consistency checking for sentiment dictionaries and inconsistencies within individual dictionaries and across dictionaries can be pinpointed by our techniques.

8.CONCLUSION

We study the problem of checking polarity consistency for sentiment word dictionaries. We prove that this problem is NP-complete. In practice polarity inconsistencies of words both within a dictionary and across dictionaries can be obtained using SAT solvers. We study the problem of checking polarity consistency for sentiment word dictionaries. We prove that this problem is NP-complete. We show that in practice polarity inconsistencies of words both within a dictionary and across dictionaries can be obtained using SAT solvers. Sets of inconsistent words are pinpointed and this allows the dictionaries to be improved. Experiments with five sentiment dictionaries, including the union dictionary, are reported. There are several directions we plan to pursues in the future. First, we plan to categorize the polarity inconsistencies according to our classification (Section 3) and identify the reason behind each inconsistency. Second, as more and more polarity inconsistencies will be “repaired” we will analyze the correlation rate between polarity inconsistency in a dictionary and its effect on the results in sentiment analysis tasks.

induction: corpora- and WordNet-based. Our approach falls into

ISSN : 2454-9924

REFERENCES

[1] B. Pang and L. Lee, "A sentimental education: Sentiment anal-ysis using subjectivity summarization based on minimum cuts," in Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics, 2004, pp. 271–278.

[2] C. Danescu-N.-M., G. Kossinets, J. Kleinberg, and L. Lee, "How opinions are received by online communities: A case study on amazon.com helpfulness votes," in Proc. 18th Int. Conf. World Wide Web, 2009, pp. 141–150.

[3] M. Kim and E. Hovy, "Determining the sentiment of opinions," in Proc. 20th Int. Conf. Comput. Linguistics, 2004, pp. 1367–1373.

[4] H. Takamura, T. Inui, and M. Okumura, "Extracting semantic ori-entations of words using spin model," in Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics, 2005, pp. 133–140.

[5] E. Breck, Y. Choi, and C. Cardie, "Identifying expressions of opin-ion in context," in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, pp. 2683–2688.

[6] X. Ding and B. Liu, "Resolving object and attribute coreference in opinion mining," in Proc. 23rd Int. Conf. Comput. Linguistics, 2010, pp. 268–276.

[7] A. L. Maas, R. E. Daly, P. Pham, D. Huang, A. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, 2011, pp. 142–150.

[8] P. Stone, D. Dunphy, M. Smith, and J. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA, USA: MIT Press, 1996.

[9] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proc. Conf. Human Language Technol. Empirical Methods Natural Language Process., 2005, pp. 347–354.

[10] M. Taboada and J. Grieve, "Analyzing appraisal automatically," in Proc. AAAI Spring Symp., 2004, pp. 158–161.

[11] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion min-ing," presented at the 7th Int. Conf. Language Resources and Evaluation, Valletta, Malta, May 2010.

[12] R. Agerri and A. Garc_ıa-Serrano, "Q-wordnet: Extracting polarity from wordnet senses," presented at the 7th Int. Conf. Language Resources and Evaluation, Valletta, Malta, 2010.

[13] S. A. Cook, "The complexity of theorem-proving procedures," in Proc. 3rd Annu. ACM Symp. Theory Comput., 1971, pp. 151–158.

[14] A. Biere, A. Biere, M. Heule, H. van Maaren, and T. Walsh, Hand-book of Satisfiability: Volume 185 Frontiers in Artificial Intelligence and Applications. Amsterdam, The Netherlands: IOS Press, 2009.

[15] E. Dragut, H. Wang, C. Yu, P. Sistla, and W. Meng, "Polarity con-sistency checking for sentiment dictionaries," in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics: Long Papers, 2012, pp. 997–1005.

[16] J. Han, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2005.

[17] S.-M. Kim and E. Hovy, "Identifying and analyzing judgment opinions," in Proc. Main Conf. Human Language Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2006, pp. 200–207.

[18] A. Andreevskaia and S. Bergler, "Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses," in Proc. 11th Conf. Eur. Chapter Assoc. Comput. Linguistics, 2006, pp. 209–216.

[19] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta, "Revising the wordnet domains hierarchy: Semantics, coverage and bal-ancing," in Proc. Workshop Multilingual Linguistic Resources, 2004, pp. 101–108.

[20] L. Xu, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Satzilla: Port-folio-based algorithm selection for SAT," J. Artif. Int. Res., vol. 32, pp. 565–606, 2008.

[21] D. Babic, J. Bingham, and A. J. Hu, "B-cubing: New possibilities for efficient sat-solving," IEEE Trans. Comput., vol. 55, no. 11, pp. 1315–1324, Nov. 2006.

[22] P. Jackson and D. Sheridan, "Clause form conversions for boolean circuits," in Proc. 7th Int. Conf. Theory Appl. Satisfiability Testing, 2004, pp. 183–198.

[23] V. P. Nelson, H. T. Nagle, B. D. Carroll, and J. D. Irwin, Digital Logic Circuit Analysis and Design. Englewood Cliffs, NJ, USA: Pren-tice-Hall, 1995.

[24] N. Dershowitz, Z. Hanna, and E. Nadel, "A scalable algorithm for minimal unsatisfiable core extraction," in Proc. 9th Int. Conf. Theory Appl. Satisfiability Testing, 2006, pp. 36–41.

[25] D. L. Berre and A. Parrain, "The sat4j library, release 2.2," J. Satis-ability, Boolean Model. Comput., vol. 7, no. 2/3, pp. 59–64, 2010.

[26] A. Biere, "Picosat essentials," J. Satisability, Boolean Model. Comput., vol. 4, no. 2–4, pp. 75–97, 2008.

[27] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool, 2012.

[28] J. Kamps, M. Marx, R. Mokken, and M. de Rijke, "Using wordnet to measure semantic orientation of adjectives," in Proc. 4th Int. Conf. Language Resources Eval., 2004, pp. 1115–1118.

[29] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," in Proc. 11th Conf. Eur. Chap-ter Assoc. Comput. Linguistics, 2006, pp. 193–200.

ISSN : 2454-9924

[30] D. Rao and D. Ravichandran, "Semi-supervised polarity lexicon induction," in Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguis-tics, 2009, pp. 675–682.

[31] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in Proc. 14th ACM Int. Conf. Inform. Knowl. Manage., 2005, pp. 617–624.

[32] A. Hassan and D. Radev, "Identifying text polarity using random walks," in Proc. 48th Annu. Meeting Assoc. Comput. Linguistics, 2010, pp. 395–403.

[33] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," ACM Trans. Inform. Syst., vol. 21, pp. 315–346, 2003.

[34] C. Akkaya, J. Wiebe, and R. Mihalcea, "Subjectivity word sense disambiguation," in Proc. Conf. Empirical Methods Natural Language Process., 2009, pp. 190–199.

Eduard C. Dragut received the PhD degree in computer science from the University of Illinois at Chicago in 2010. He is currently an assistant professor in the Department of Computer and Information Sciences, Temple University. His research interests include web-based information retrieval and web database integration. He is a coauthor of the book Deep Web Query Interface Understanding and Integration. He co-chaired the VLDB QDB 2012 and ICDE PhD Sympoisum 2014 workshops. He is a member of the IEEE.

Hong Wang is working toward the PhD degree in the Department of Computer Science, University of Illinois at Chicago. He has interned with Goo-gle, and with Huawei Technologies as a software engineer. His research interests include web table understanding, information extraction, and natural language processing.

Prasad Sistla received the PhD degree in com-puter science/applied mathematics from Harvard University in 1983. He is currently a professor in the Department of Computer Science, University of Illinois at Chicago. He has done extensive research in model checking, formal methods, databases, and security. He has published exten-sively in leading computer science venues. He cochaired and served on the Program Commit-tees of leading Computer Science conferences. His research has been funded by leading organi-zations such as US National Science Foundation (NSF), AFOSR, and US Defense Advanced Research Projects Agency (DARPA).

Clement Yu received the PhD degree in com-puter science from Cornell University in 1973. He is a professor in the Department of Computer Sci-ence, University of Illinois at Chicago. His areas of interest include search engines and multimedia retrieval. He has published in leading journals such as the IEEE TKDE, ACM TODS, and Jour-nal of the ACM and conferences such as VLDB, ACM SIGMOD, ACM SIGIR. He served as the chairman of the ACM SIGIR, the general chair of ACM SIGMOD, and as a member of the advisory committee to the US National Science Foundation (NSF). He was/is a member of the editorial boards of IEEE TKDE, IJSEKE, Distributed and Parallel Databases, and WWW Journal.

Weiyi Meng received the PhD degree in com-puter science from the University of Illinois at Chi-cago in 1992. He is currently a professor in the Department of Computer Science, State Univer-sity of New York at Binghamton. His research interests include web-based information retrieval, metasearch engines, and web database integra-tion. He is a coauthor of three books Principles of Database Query Processing for Advanced Appli-cations, Advanced Metasearch Technology, and Deep Web Query Interface Understanding andIntegration. He is on the editorial boards of WWW Journal and Frontiers of Computer Science. He is a member of the IEEE." For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dli