POLARITY CONSISTENCY CHECKING FOR MULTIPOLARITY BASED DOMAIN INDEPENDENT SENTIMENT DICTIONARIES

C.SUJITHRA AND A.ARUNKUMAR
PG Scholar, Assistant Professor,
Department of Computer Science and Engineering,
Sri Krishna College of Engineering and Technology
15epcs017@skcet.ac.in, aarunkumar.83@gmail.com

Abstract

Polarity classification is very important for sentiment analysis. Polarity consistency checking problem is used to find all the polarity inconsistencies word in sentiment dictionary. We perform experiments on four sentiment dictionaries and WordNet. Several domain independent sentiment dictionaries have been manually or semi automatically created. OF, GI and AL are called sentiment word dictionaries (SWD). The domain dependent dictionaries are constructed by using the positive negative terms based on the particular domain. To construct these dictionaries, the domain knowledge plays a vital role. We proposed a new approach in which we implement hypernym with WordNet. Hyponyms are subdivisions of more general words. The semantic relationship between each of the more specific words (e.g., daisy and rose) and the more general term (flower) is called hyponymy or inclusion. In this system we include subjective and objective senses of a word. It is implemented by using Subjectivity Word Sense Disambiguation. Finally in this proposed system polarity distribution is implemented with multiple polarities. It increases the accuracy of the polarity consistency check. We reduce the polarity consistency problem to the satisfiability problem and utilize two fast SAT solvers to detect inconsistencies in a sentiment dictionary. We perform experiments on five sentiment dictionaries and Wordnet to show inter- and intra-dictionaries inconsistencies.

Keywords- WordNet, SWD, PCC, Sentiment dictionary.

I. INTRODUCTION

Opinion mining can be useful in several ways. It can help marketers evaluate the success of an ad campaign or new product launch, determine which versions of a product or service are popular and identify which demographics like or dislike particular product features. For example, a review on a website might be broadly positive about a digital camera [2], but be specifically negative about how heavy it is. Being able to identify this kind of information in a systematic way gives the vendor a much clearer picture of public opinion than surveys or focus groups do, because the data is created by the customer. There are different methods are used for opinion mining. Some of them are given below:

- Subjectivity/objectivity identification
- Feature/aspect-based

Subjectivity/objectivity identification is commonly defined as classifying a given text into one of two classes: objective or subjective. This problem can sometimes be more difficult than polarity classification. The subjectivity words [34] depend on context and an objective may contain subjective sentence (e.g., a news article quoting). Moreover, results are largely dependent on the definition of subjectivity used when annotating texts. Feature/aspect-based refers to determining the opinions or sentiments expressed on different features or aspects of entities, e.g., of a cell phone, a digital camera, or a bank. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone, the service for a restaurant, or the picture quality of a camera. The advantage of feature-based sentiment analysis is the possibility to capture nuances about objects of interest. Different features can generate different sentiment responses, for example a hotel can have a convenient location, but mediocre food. This problem involves several sub-problems, e.g., identifying relevant entities, extracting their features/aspects, and determining whether an opinion expressed on each feature/aspect is positive, negative or neutral. The automatic identification of features can be performed with syntactic methods or with topic modeling. More detailed discussions about this level of sentiment analysis can be found in Liu's work [27].Polarity consistency checking is finding inconsistencies problem in a sentiment word dictionary. We give a solution that reduces an instance of the problem to an instance of CNF-SAT. We can apply the fast SAT solver problem to solve our problem [20]. Boolean logic, a formula is in conjunctive normal form (CNF) or clausal normal. Otherwise put, it is an AND of ORs. All conjunctions of literals and all disjunctions of literals are in CNF, as they can be seen as conjunctions of one-literal clauses and conjunctions of a single clause, respectively. As in the disjunctive normal form (DNF), the only propositional connectives a formula in CNF can contain are AND or OR. In automated theorem proving, the notion "clausal normal form" is often used in a narrower sense, meaning a particular representation of a CNF formula as a set of sets of literals. Boolean formula [22] assigns variables true or false.

Polarity inconsistency problem is

- Sentiment dictionaries do not address the concept of polarity inconsistency of a words/synsets.
- We define consistency among the polarities of words/synsets in a dictionary and give methods to check it.
- Hence, its conveys a positive sentiment is when used with this sense.
- Manual checking of sentiment dictionaries for inconsistency is a difficult endeavor.
- We aim to unearth these inconsistencies in sentiment dictionaries.
- The presence of inconsistencies found via polarity analysis is not exclusively attributed to one party.
- Therefore, a by-product of our polarity consistency analysis is that it can also locate some of the likely places where WordNet needs linguists' attention.

The domain independent dictionaries are constructed by using the general positive negative term. To construct these dictionaries, the domain knowledge is not necessary. The domain dependent dictionaries are constructed by using the positive negative terms based on the particular domain. To construct these dictionaries, the domain knowledge plays a vital role.

II. RELATED WORK

There are two lines of work on sentiment polarity lexicon induction: corpora- and WordNet-based. Our approach falls into the latter. WordNet-based approaches use lexical relations defined in WordNet to derive sentiment lexicons. For example, [28] determines sentiments of adjectives in WordNet by measuring the relative distance of a term from exemplars, such as“good” and “bad”. The work reports results for adjectives alone. Other approaches use synonyms and antonyms to expand the sets of seeds [29]. Yet another technique is to add all synonyms of a polar word with the same polarity and its antonyms with reverse polarity [17]. Moreover, we have encountered instances of antonym pairs where the polarity is not necessarily reversed (e.g., the adjective advance has a positive polarity while one of its antonyms, middle, has neutral polarity). QW [12] 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. Also, they do not assign polarities to words. Finally, the relations in Word- Net do not have well-defined behavior with respect to preserving/ reversing polarity. Instead, each synset is 100 percent positive, 100 percent negative or 100 percent neutral. Machine learning algorithms [31] as well as stochastic algorithms [32] can be employed to classify words into different polarities. According to [29], the performance of [31] is comparable or better than those in [28], [33]. The differences between our approach and earlier ones including those that are not WordNet based, are: (1) to our knowledge, none of the earlier works studied the problem of polarity consistency checking for sentiment dictionaries and [15] inconsistencies within individual dictionaries and across dictionaries can be pinpointed by our techniques

III. PROPOSED ARCHITECTURE

The opinions expressed in various web and media outlets (e.g., blogs, newspapers) are an important yardstick 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 sentence/ document is by analysis of the orientations of the individual words. Sentiment dictionaries [35] are utilized to facilitate the summarization.

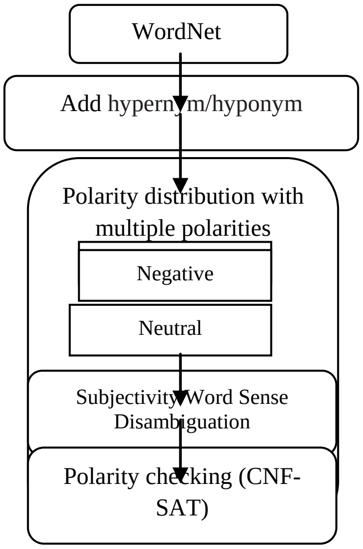
WordNet
Add hypernym/hyponym
Polarity distribution with multiple polarities
Negative
Neutral
Subjectivity Word Sense Disambiguation
Polarity checking (CNF-SAT)

Fig 1: Proposed Architecture

A dictionary is a collection of words in one or more specific languages, often listed alphabetically, with usage information, definitions, [24] etymologies, phonetics, pronunciations, and other information; or a book of words in one language with their equivalents in another, also known as a lexicon.

Word sense [18] in linguistics, a word sense is one of the meanings of a word. For example a dictionary may have over 50 different meanings of the word play, each of these having a different meaning based on the context of the word usage in a sentence. For example: We went to see the play Romeo and Juliet at the theater. The children went out to play in the park. In each sentence we associate a different meaning of the word "play" based on hints the rest of the sentence gives us. WordNet is a lexical database for the English language. It groups English words into sets of synonyms[15] called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis[17] and artificial intelligence applications. Sentiment analysis or opinion mining refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information[4] in source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. NP complete in computational complexity theory, the complexity class NP-complete (abbreviated NP-C or NPC) is a class of decision problems. A decision problem L is NP-complete if it is in the set of NP problems so that any given solution to the decision problem can be verified in polynomial time, and also in the set of NP-hard problems so that any NP problem can be converted into L by a transformation of the inputs in polynomial time.

IV. WORDNET

Opinion mining is a lexical resource for SentiWordNet. It assigns to each sense of WordNet, a lexical resource consisting of WordNet senses automatically annotated by positive and negative polarity. Polarity classification amounts to decide whether a text may be associated to positive or negative connotations. Polarity classification is becoming important for applications such as Sentiment Analysis, which facilitates the extraction and analysis of opinions about commercial products, to track attitudes by mining online forums, blogs, on companies reputation management, brand monitoring etc. Inspired by work on classification of word senses by polarity, and taking WordNet as a starting point, we build WordNet[10]. Instead of applying external tools such as supervised classifiers to annotated WordNet sense by polarity, we try to effectively maximize the linguistic information contained in WordNet, advantage of the human effort put by lexicographers and annotators. WordNet consists of words, synsets and frequency counts. This is a standard smoothing technique [16].

V. INCONSISTENCY CLASSIFICATION

Input Dictionaries Polarity Inconsistency

Input polarity inconsistencies are of two types: intra-dictionary and inter-dictionary inconsistencies[35]. The latter are obtained by comparing two Sentiment word dictionaries, a Sentiment word dictionary with an sentiment sense dictionary and two sentiment sense dictionary.

Intra-Dictionary Inconsistency

A Sentiment word dictionary may have triplets of the form positive, negative and neutral For instance, the verb brag has both positive and negative polarities in OpionionFinder[6]. For these cases, we apply to determine the polarity of word w with part of speech pos. The verb brag has negative polarity. Such cases simply say that the team who constructs the dictionary believes brag has multiple polarities as they do not adopt our dominant sense principle. Wordnet [28], a sentiment sense dictionary, does not have inconsistencies as it does not have a synset with multiple polarities.

Inter-Dictionary Inconsistency

A word belongs to this category if it appears with different polarities in different SWDs. For instance, the adjective joyless has positive polarity in Opinion Finder[7] and negative polarity in General inquirer[18]. For example depicts the overlapping relationships between the three Sentiment word dictionarys: OpionionFinder has 2,933 words in common with General Inquirer. The three dictionaries largely agree on the polarities of the words they pair wise share. For instance, out of 2,924 words shared by OpionionFinder and General Inquirer, [8] 2,834 have the same polarities. However, there are also a significant number of words which have different polarities across dictionaries. OpionionFinder and General Inquirer disagree on the polarities of 90 words. Among the three dictionaries there are 181 polarity inconsistent words. The polarities of these words are manually corrected using before the polarity consistency checking is applied to the union of the three dictionaries.

VI. SUBJECTIVITY WORD SENSE DISAMBIGUATION

Subjectivity word sense disambiguation is midway between pure dictionary classification and pure contextual interpretation. For, Subjectivity word sense disambiguation [34] the context of the word is considered in order to perform the task, but the subjectivity is determined solely by the dictionary. In contrast, full contextual interpretation can deviate from a sense's subjectivity label in the dictionary. As noted above, words used with objective senses may appear in subjective expressions. We use a supervised approach to Subjectivity [34]word sense disambiguation. We train a different classifier for each lexicon entry for which we have training data. The training and test data for Subjectivity word sense disambiguation consists of word instances in a corpus labeled as Subjective or Objective, indicating whether they are used with a subjective or objective sense. Many approaches to opinion, sentiment, and Subjective analysis on lexicons of words that may be used to express subjectivity. Examples of such words are the following (in bold):

> (1) He is a **disease** to every team he has gone to.
> Converting to SMF is a **headache**.
> The concert left me **cold**.
> That guy is such a **pain**.

Knowing the meaning and thus subjectivity of
Disease, headache, cold, pain these words would help a system recognize the negative sentiments in these sentences. Most subjective lexicons are compiled as lists of keywords, rather than word meanings senses. However, many words have both subjective and objective senses. False hits subjectivity clues used with objective senses are a significant source of error in subjectivity and sentiment analysis. For example, even though the following sentence contains all of the negative words above, it is nevertheless objective, as they are all false hits:

> (2) Early symptoms of the **disease** include severe **headaches**, red eyes, fevers and **cold** chills, body **pain**, and vomiting.

We define a new task, *subjectivity word sense disambiguation*,[34] which is to automatically determine which word instances in a corpus are being used with subjective senses, and which are being used with objective senses.

VII. COMPLEX POLARITY INCONSISTENCY

This kind of inconsistency is more subtle and cannot be detected by direct comparison of words/synsets. They consist of sets of words and/or synsets whose polarities [17] cannot concomitantly be satisfied. By assuming that WordNet is correct, it is not possible for the two words to have different polarities: the sole synset, which they share, would have two different polarities, which is a contradiction. The occurrence of an inconsistency points out the presence of incorrect input data: The information given in WordNet is incorrect or the information in the given sentiment dictionary [28]is incorrect, or both.

Polarity Consistency Checking

To "exhaustively" solve the problem of finding the polarity inconsistencies in an Sentiword dictionary, we propose a solution that reduces an instance of the problem to an instance of CNF-SAT. We can then employ a fast SAT solver [25] to solve our problem. CNF-SAT is a decision problem of determining if there is an assignment of True and False to the variables of a Boolean formula Φ in conjunctive normal form (CNF) such that Φ evaluates to True[20], [21]. A formula is in CNF if it is a conjunction of one or more clauses, each of which is a disjunction of literals. CNF-SAT is a classic NP-complete problem, but, modern SAT solvers are capable of solving many practical instances of the problem. Since, in general, there is no easy way to tell the difficulty of a problem without trying it, SAT solvers [25] include time-outs, so they will terminate even if they cannot find a solution.

There are two methods of converting an instance of the polarity consistency checking problem into an instance of CNF-SAT. The first method, called exhaustive enumeration of MDS method (EEM), the second method, called frequency summation method (FSM).

Exhaustive Enumeration of MDSs Method (EEM)

We now elaborate the construction of C. We enumerate all the MDSs of w and for each of them we introduce a clause. The clauses are then concatenated by OR in the Boolean formula. The formula Φ is not in CNF after this construction and it needs to be converted. The conversion to CNF [21]is a standard procedure and we omit it in this paper. Φ in CNF is input to a SAT solver.

Frequency Summation Method (FSM)

This method used for reducing an instance of the PCC problem into an instance of CNF-SAT, which gives a polynomial length formula for C(w, p). The idea is to simulate a logic circuit that evaluates Inequality 3 and outputs true when this inequality is satisfied and false when it is not. Then, we derive the Boolean expression associated with the circuit. A careful analysis of the inequality reveals that we need three main circuit components: a SUM component that computes the summation $\sum_{s\in S_w} f(w,s)pol(s,p)$an Instantiation component that evaluates each term $f(w,s)pol(s,p)$ before it is input to the SUM component and a Digital Comparator component that asserts the inequality. Bottom up, the logic circuit is constructed as follows:

1. For each $s \in S_w$ we need an Instantiation component Is. The inputs of Is are f(w,s) and s_p. Is outputs f(w, s) if s_p = True (i.e., the synset s has polarity p) and outputs 0 if s_p = False (i.e., s does not have polarity p).
2. The SUM [23] component adds the outputs of Is's pair wise; then, it adds their results pair wise; so on. This scheme can be captured as a full binary tree whose leaf nodes denote the frequencies of use of the synsets and whose internal nodes represent the sum of values of the frequencies.
3. The output of SUM is input together with the constant $\frac{1}{2}freq(w)$ to the Digital Comparator.

VIII. CONCLUSION

We study the problem of checking polarity consistency for sentiment word dictionaries. We include hypernym with WordNet. Hyponyms are subdivisions of more general words. We include subjective and objective senses of a word. It is implemented by using Subjectivity Word Sense Disambiguation. It automatically determines which word instances in a corpus are being used with subjective senses, and which are being used with objective senses. Finally polarity distribution is implemented with multiple polarities. It increases the accuracy of the polarity consistency check. We reduce the polarity consistency problem to the satisfiability problem and utilize two fast SAT solvers to detect inconsistencies in a sentiment dictionary. A set of inconsistent words allows the dictionaries to be improved. We performed experiments on five sentiment dictionaries and WordNet to show inter- and intra-dictionaries inconsistencies.

IX. REFERENCES

[1] B. Pang and L. Lee, “A sentimental education: Sentiment analysis 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 orientations 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 opinion 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 mining,” 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, Handbook 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 consistency 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 balancing,” in Proc. Workshop Multilingual Linguistic Resources, 2004, pp. 101–108.

[20] L. Xu, F. Hutter, H. H. Hoos, and K. Leyton-Brown, “Satzilla: Portfolio- 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: Prentice- Hall, 1995.

[24] N. Dershowitz, Z. Hanna, and E. Nadel, “A scalable algorithm forminimal 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. Satisability, 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. Chapter Assoc. Comput. Linguistics, 2006, pp. 193–200.

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

[31] A. Esuli and F. Sebastiani, “Determining the semantic orientationof 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. 19.

[35] Eduard C. Dragut, Hong Wang, Prasad Sistla, Clement Yu, and Weiyi Meng,” Polarity Consistency Checking for Domain Independent Sentiment Dictionaries” IEEE transactions on knowledge and data engineering, vol. 27, no. 3, march 2015.