Information Processing and Management 57 (2020) 102273

Contents lists available at ScienceDirect

Information Processing and Management

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

Deriving the sentiment polarity of term senses using dual-step context-aware in-gloss matching

Mohammad Darwich, Shahrul Azman Mohd Noah*, Nazlia Omar

Centre for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Malaysia

ARTICLE INFO

Keywords:
Sentiment lexicon
Opinion lexicon
Sentiment lexicon generation
Sentiment analysis
Opinion mining

ABSTRACT

Vital to the task of Sentiment Analysis (SA), or automatically mining sentiment expression from text, is a sentiment lexicon. This fundamental lexical resource comprises the smallest sentiment-carrying units of text, *words*, annotated for their sentiment properties, and aids in SA tasks on larger pieces of text. Unfortunately, digital dictionaries do not readily include information on the sentiment properties of their entries, and manually compiling sentiment lexicons is tedious in terms of annotator time and effort. This has resulted in the emergence of a large number of research works concentrated on automated sentiment lexicon generation. The dictionary-based approach involves leveraging digital dictionaries, while the corpus-based approach involves exploiting co-occurrence statistics embedded in text corpora. Although the former approach has been exhaustively investigated, the majority of works focus on terms. The few state-of-the-art models concentrated on the finer-grained term sense level remain to exhibit several prominent limitations, e.g., the proposed semantic relations algorithm retrieves only senses that are at a close proximity to the seed senses in the semantic network, thus prohibiting the retrieval of remote sentiment-carrying senses beyond the reach of the 'radius' defined by number of iterations of semantic relations expansion. The proposed model aims to overcome the issues inherent in dictionary-based sense-level sentiment lexicon generation models using: (1) null seed sets, and a morphological approach inspired by the Marking Theory in Linguistics to populate them automatically; (2) a dual-step context-aware gloss expansion algorithm that 'mines' human defined gloss information from a digital dictionary, ensuring senses overlooked by the semantic relations expansion algorithm are identified; and (3) a fully-unsupervised sentiment categorization algorithm on the basis of the Network Theory. The results demonstrate that context-aware in-gloss matching successfully retrieves senses beyond the reach of the semantic relations expansion algorithm used by prominent, well-known models. Evaluation of the proposed model to accurately assign senses with polarity demonstrates that it is on par with state-of-the-art models against the same gold standard benchmarks. The model has theoretical implications in future work to effectively exploit the readily-available human-defined gloss information in a digital dictionary, in the task of assigning polarity to term senses. Extrinsic evaluation in a real-world sentiment classification task on multiple publically-available varying-domain datasets demonstrates its practical implication and application in sentiment analysis, as well as in other related fields such as information science, opinion retrieval and computational linguistics.

* Corresponding author.
E-mail address: shahrul@ukm.edu.my (S.A.M. Noah).

https://doi.org/10.1016/j.ipm.2020.102273
Received 19 July 2019; Received in revised form 12 April 2020; Accepted 18 April 2020
Available online 05 June 2020
0306-4573/

1. Introduction

Sentiment analysis (SA) is at the crossroads of natural language processing and information science, and involves the detection of sentiment, attitude, opinion and emotion in unstructured, free-form text (Poria, Majumder, Mihalcea & Hovy, 2019; Hussein, 2018; Chaturvedi, Cambria, Welsch, & Herrera, 2018; Mäntylä, Graziotin, Kuutila, 2018; Liu, 2017; Al-Saffar, Awang, Tao, Omar, Al-Saiagh, & Al-bared, 2018). The (unsupervised) lexicon-based approach involves making use of a sentiment lexicon to compute the global sentiment polarity of a text document, based on the aggregation of the polarity of the individual words embedded within the document (Alqasemi et al., 2019; Saif et al., 2017; Fernández-Gavilanes, Álvarez-López, Juncal-Martínez, Costa-Montenegro, & González-Castaño, 2016; Hutto, & Gilbert, 2014; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Conversely, the (supervised) classification-based approach involves constructing supervised machine learning classifiers that are fed with manually labeled training data for the classification task (Yang, Zhang, Jiang, & Li, 2019; Xing, Pallucchini, & Cambria, 2019; Thelwall, 2017).

Supervised machine learning classifiers are able to optimise informative features on large-scale datasets, a task beyond the reach of human capability and introspection. However, a reoccurring issue inherent in employing classifiers is that they rely on large manually-labelled datasets, most of which must be tailored to specific domains. They are also are computation-intensive, and involve a hidden reasoning process due to their 'black-box' nature. An alternative choice is to take an unsupervised, lexicon-based approach.

The limitations of document-level sentiment analysis based on predefined sentiment lexicons are well-known (Hamilton, Clark, Leskovec, & Jurafsky, 2016; Li, & Hovy, 2017; Liu, 2015, 2017). For example, the sentiment of a given piece of text may differ depending on who is the author or who is reading that text (Hamilton et al., 2016; Liu, 2015). Despite these limitations, lexicon-based approaches heavily rely on the lexicon itself as the main source of input, and thus do not require any 'human training and supervision', such as manually-labelled training datasets. Moreover, they are generally robust in variable domain scenarios (as empirically demonstrated by Taboada et al. (2011) and Hutto and Gilbert (2014)), computationally economical in real-time processing, and relatively easy to modify due to their 'transparent' nature.

Modern SA and opinion retrieval systems in information science typically assign a text unit to one of a set of predefined classes (positive, negative or objective), at the document, sentence, subsentence/clause, phrase or aspect level. Prior to performing SA on larger pieces of text, however, there is an essential need to classify the smallest sentiment-carrying lexical units of text, *words*, with their corresponding sentiment properties. The outcome is a sentiment lexicon, which is an essential resource that drives the performance of SA models, and equips them with the ability to effectively determine the global sentiment of larger pieces of text. Unfortunately, digital dictionaries such as the Princeton WordNet and Merriam Webster do not readily include information on the sentiment properties of their entries. Moreover, manually assigning entry terms with their corresponding sentiment properties is prohibitively onerous in terms of annotator time and effort (Schneider, Male, Bhogadhi, & Dragut, 2018; Dragut, Wang, Yu, Sistla, & Meng, 2012). The explosive research interest in the field of SA during the past few years, along with the restrictions of manual sentiment lexicon compilation, has resulted in a high demand for an automatic means of generating reliable sentiment lexicons, which is beyond the means of manual human annotation.

Automated sentiment lexicon generation branches in two general directions. The dictionary-based approach involves leveraging digital dictionaries to tag terms with their corresponding sentiment properties (e.g. San Vicente, Agerri, & Rigau, 2014; Baccianella, Esuli, & Sebastiani, 2010; Hassan, & Radev, 2010). Conversely, the corpus-based approach involves exploiting co-occurrence statistics or syntactic patterns in text corpora (e.g. Alqasemi, Abdelwahab, & Abdelkader, 2019; Peng, & Park, 2011). The underlying intuition is that entries in an online dictionary are not only semantically related by meaning, but for the most part, are also related in terms of their sentiment properties.

The majority of sentiment lexicon generation models in the literature focus on *terms* (Alqasemi, Abdelwahab, & Abdelkader, 2019; Peng, & Park, 2011; Hassan, & Radev, 2010; Rao, & Ravichandran, 2009; Blair-Goldensohn, Hannan, McDonald, Neylon, Reis, & Reynar, 2008; Williams, & Anand, 2009). However, the models in these works are not sensitive enough to detect the alteration in the sentiment properties of the various senses of polysemous terms. As a solution, other works focus on *term senses* (San Vicente et al., 2014; Agerri, & Garcia-Serrano, 2010; Baccianella et al., 2010). In this case, each unique sense is treated individually, since the various senses of a term may have different meanings, and in turn different sentiment properties (Akkaya, Wiebe, & Mihalcea, 2009; Wiebe, & Mihalcea, 2006).

According to the literature, several problems inherent in existing dictionary-based sense-level lexicon generation models restrict their ability to assign sentiment polarity to term senses with optimal performance. First, they all require manually-labeled seed sets as input. Second, the creators of the popular and widely-applied SentiWordNet (Baccianella et al., 2010) employ a semantic relations algorithm that only retrieves senses that are at a close proximity to the seed senses, thus overlooking remote potentially-subjective senses in the semantic network. Third, some require supervised classifiers that prioritize the objective class (Baccianella et al., 2010) (), while others treat objectivity as non-existent entity and aggressively filter out non-polar senses (San Vicente et al., 2014; Agerri, & Garcia-Serrano, 2010), resulting in the labeling of subtly subjective senses as objective.

To overcome these prominent limitations, the *primary objective* of this work is to propose a sense-level sentiment lexicon generation model that:

- initiates with null seed sets, and a morphology-inspired approach based on the Marking Theory (Battistella, 1990; Lehrer, 1974) in Linguistics to populate them automatically;
- employs a dual-step context-aware in-gloss matching algorithm to detect remote sentiment-carrying senses in the semantic network; and
- utilizes a fully-unsupervised sentiment categorization algorithm based on the Network Theory (Burt, 1980).

Table 1
Sentiment lexicon generation models for various languages.

Language	References
Arabic	(Al-Ayyoub et al., 2019; Alshahrani, & Fong, 2018; Al-Thubaity, Alqahtani, & Aljandal, 2018; Assiri, Emam, & Al-Dossari, 2018; Guellil et al., 2018; El-Beltagy, 2016; Mohammad, Salameh, & Kiritchenko, 2016)
Turkish	(Ekinci, & Omurca, 2019; Dehkharghani, 2018)
Chinese	(Zhang et al., 2018; Kong et al., 2018; Wu et al., 2016)
Russian	(Loukachevitch, & Levchik, 2016; Koltsova, Alexeeva, & Kolcov, 2016)
French	(Kandé et al., 2019; Abdaoui et al., 2017)
Swedish	(Rouces et al., 2018)
Portuguese	(Machado, Pardo, & Ruiz, 2018)
Indonesian	(Saputra, & Nurhadryani, 2018)
Malay	(Darwich, Noah, & Omar, 2017, 2016, 2015)
Thai	(Suktarachan, 2018)
Urdu	(Asghar et al., 2019; Rehman, & Bajwa, 2016)
Tamil	(Kannan, 2019)
Vietnamese	(Tran, & Phan, 2018)
Korean	(Song, Park, & Shin, 2019)

Intrinsic evaluation of the model to accurately assign senses with polarity demonstrates that it is on par with human judgement, as well as with state-of-the-art models, against the same gold standard benchmarks. Extrinsic evaluation of the resultant lexicon in a real-world sentiment classification task on four publically-available varying-domain datasets demonstrates its superior performance compared to that of related sense-level lexicons, as well as the improved effectiveness it has over a 'coarse' term-level version of the same lexicon.

The remainder of this paper is outlined as follows. Section 2 presents prior work on sentiment lexicon generation. Section 3 discusses the methods carried out to develop the proposed model. Section 4 discusses the experimental setup and evaluation procedure. Section 5 presents the results and discussion. Section 6 concludes, and highlights some worthy recommendations for future work.

2. Prior work

Numerous works in the related literature have been devoted to automatic generation of sentiment lexicons for the English language (Alqasemi, Abdelwahab, & Abdelkader 2019; Saif et al., 2017; San Vicente, Agerri, & Rigau, 2014; Baccianella et al., 2010). This has also been attempted for other languages, as shown in Table 1. Generalized algorithms that are able to generate lexicons for multiple languages have also been developed (Kaity, & Balakrishnan, 2018, 2019; Asgari, Braune, Ringlstetter, & Mofrad, 2019; San Vicente et al., 2014).

The dictionary-based approach involves leveraging lexical resources and online dictionaries (WordNet, Merriam Webster, etc.) to automatically tag terms with their corresponding sentiment polarity (e.g. Baccianella et al., 2010; Hassan, & Radev, 2010). Conversely, the corpus-based approach involves exploiting co-occurrence statistics or syntactic patterns in a text corpus (e.g. Alqasemi et al., 2019; Saif, Fernandez, Kastler, & Alani, 2017; Deng, Sinha, & Zhao, 2017; Fernández-Gavilanes et al., 2016; Peng, & Park, 2011; Osman, Noah, & Darwich, 2019). The end result of both approaches is a *sentiment lexicon*, or a list of natural language vocabulary terms marked with their underlying sentiment properties (polarity and strength).

Text corpora have been commonly used in domain adaptation, which involves converting a domain-independent sentiment lexicon into a domain-specific lexicon, or domain specific lexicon into an entirely different domain (Alqasemi, Abdelwahab, & Abdelkader, 2019; Saif et al., 2017; Deng et al., 2017; Fernández-Gavilanes et al., 2016; Weichselbraun, Gindl, & Scharl, 2011; Bollegala, Weir, & Carroll, 2011). Semi-supervised label propagation has been heavily investigated (Wang, Pan, Dahlmeier, & Xiao, 2017; Hamilton et al., 2016; Huang, Niu, & Shi, 2014; Tai, & Kao, 2013; Velikovich et al., 2010). For example, Hamilton et al. (2016) utilize a text corpus and combine label propagation and word vector embeddings to induce domain-specific sentiment lexicons using small predefined seed sets.

Social media corpora such as Twitter have been utilized to generate informal social media-specific lexicons (Wu, Wu, Chang, Wu, & Huang, 2019; Wu, Morstatter, & Liu, 2018; Kimura, & Katsurai, 2017; Tang, Wei, Qin, Liu, & Zhou, 2014; Vo, & Zhang, 2016; Severyn, & Moschitti, 2015; Feng et al., 2013; Peng, & Park, 2011). Other work has investigated co-occurence metrics such as latent Dirichlet allocation (Alshahrani & Fong, 2018), as well as pointwise mutual information and normalized Google distance (Feng, Zhang, Li, Wang, Yu, & Wong, 2013; Xu, Peng, & Cheng, 2012; Taboada, Anthony, & Voll, 2006).

However, relying solely on corpora in inducing sentiment lexicons comes with inevitable limitations. First, unlike a formal dictionary, which readily comprises the entire vocabulary of a natural language, a massive corpus is required in order to capture the entire span of vocabulary words across a natural language. Second, a corpus is generally free-form and unstructured, making it 'noisy', in contrast to the structured layout of a dictionary. Third, using a corpus may also be both data- and computation-intensive. Fourth, co-occurrence statistics may not always be reliable. For example, Miller, Beckwith, Fellbaum, Gross, and Miller (1990) claim that antonyms of adjectives often co-occur together in the same phrases and sentences. Moreover, Kanayama and Nasukawa (2006) mention that only about 60% of co-occurrences reflect similar sentiment. Finally, if a corpus from only one particular domain is

available, this would adapt sentiment words to that particular domain, making them unreliable when applied in sentiment classification on an entirely different domain.

On this basis, utilizing a dictionary-based approach as an initial step, prior to the use of text corpora, is not uncommon (Peng, & Park, 2011). A dictionary in this scenario plays the role of a semantic knowledge base that typically includes extensive coverage of the entire span of vocabulary entries defined in a natural language. It possesses a neatly structured layout, in which the entry term on the left-hand and the corresponding human-defined gloss on the right-hand hold strong semantic equivalence; abides formal grammatical conventions and styles, eliminating noise that is typically found in unstructured text corpora; and exhibits an ontological network that organizes terms based on their lexical/semantic relationships (e.g., synonymy/hyponymy), whereby some relations strategically constrain the sentiment properties of terms throughout the semantic network (San Vicente et al., 2014). The dictionary-based approach has been exploited using various techniques, which are discussed hereafter.

2.1. Gloss classification

Baccianella et al. (2010) develop the popular and widely-applied sense-level sentiment lexicon referred to as SentiWordNet 3.0, whereby every synset in WordNet 2.0 is assigned three numerical values in the range of [0, 1], denoted as *Pos(s), Neg(s)* and *Obj(s)*, all of which must sum up to 1.0. The positive and negative seed sets are expanded with new synsets through semantic relations propagation, applying the semantic relations of *similarity, direct antonymy, pertains-to, derived-from, also-see* and *attribute* from the WordNet semantic network.

The glosses of the synsets in the expanded seed sets are used to generate the training data for the subsequent ternary (positive-negative-objective) classification task. The resultant features vectors from the training data are used for supervised classification, involving two binary classifiers. The first one classifies words as positive or *not positive*, and the second one classifies words as negative or *not negative*. Senses labeled as positive by the first classifier, and as *not negative* by the second classifier, are all added to the positive class. The opposite process is repeated for the negative class.

Two different classifier models (Rocchio and SVMs) were employed, each with four different training sets generated from various iterations (K) of relations propagation (K = 0, 2, 4, 6), forming the committee of eight classifiers altogether. This committee 'votes' and assigns three numerical scores (*Pos(s), Neg(s)* and *Obj(s)*) to each synset. Every label is assigned a score that is proportional to the amount of classifiers that have assigned it. A random walk step is applied after the ternary semi-supervised gloss classification step. If any synset appears in the gloss of another, they are linked, forming a graph for the random walk. The *Pos(s)* and *Neg(s)* scores are refined based on the results of the random walk step. The Princeton WordNet Gloss Corpus was employed to disambiguate terms appearing within glosses, thus taking each gloss as a 'string of synsets' rather than a string of terms.

Micro-WN(Op)-3.0 is a small portion of WordNet 3.0 (total of 1105) that was carefully manually labelled for *Pos(s), Neg(s)* and *Obj(s)* scores. The *Kendall tau distance* is a metric that counts the number of disagreements between two ranking lists. The larger this distance, the less similar the two lists are. The Kendall tau distance was employed as a metric to estimate the accuracy of the SentiWordNet 3.0 model, where, after ranking both the Micro-WN(Op)-3.0 and the corresponding SentiWordNet 3.0 synset lists by scores, any mismatches increase the tau distance, indicating poorer accuracy. The highest accuracy value achieved is 0.744.

The SentiWordNet 3.0 (Baccianella et al. 2010) generation model comes with several critical limitations. First, they manually select a total of 105 (47 positive and 58 negative) synsets as seeds. However, the entire aim in automated sentiment lexicon generation in the first place is to *minimise human involvement*. The second limitation is that a synset is labelled with both a positive score and a negative score simultaneously. They argue that each sense may be positive and negative to a certain degree, and its polarity may vary based on the context that sense is applied in. However, this is debatable. According to Lehrer (1974), "the polarity of a word is the direction it deviates from the norm from its semantic group". Therefore, assigning a polarity involves determining the degree that a term sense deviates from the norm (objectivity), and towards either positivity or negativity (and not both). This 'redundancy' is demonstrated when most SentiWordNet users simply assign a term sense with only one polarity, based on which polarity possesses the higher numerical score, or based on a subtraction of the lower score from the higher. For example, Gatti and Guerini (2012) attempt both mentioned methods as a representation of the polarity of a term using SentiWordNet.

The third limitation is that the classifiers (SVMs) used tend to emphasise a higher priority to classes with more samples, and since the expansion algorithm automatically generated the training data, there is no method to balance the volume of training data in each class. Consequently, since the objective class contained more samples than the positive and negative classes, slightly subjective words are classified as being objective.

The fourth limitation is that the use of the semantic relations expansion algorithm used to expand the initial seed sets only retrieves terms that are at a close proximity to the seed terms, as shown in Fig. 1. They choose a maximum of $K = 6$ iterations for propagation. The reason for this choice is that, the farther the distance between seeds and terms in the network (i.e. the more edges in the path that separates them), the weaker their semantic similarity becomes, and consequently, the higher the volume of 'noise' added to the seed sets. It has been well-established in the literature (Ide 2006; Wiebe & Mihalcea 2006) that the semantic significance of a path decreases as a function of its length from the source synset to the destination synset in the WordNet semantic network. Therefore, this approach prohibits the retrieval of subjective terms that are outside the reach of the radius (K) defined by the maximum number of propagation iterations. For example, most would agree that the synsets executed.a.01, evil-looking.a.01, and murky.a.02 all carry some degree of 'negativity' in reality, but are marked as objective in SentiWordNet.

The final limitation worthy to highlight is that the in-gloss matching step for the random walk is simplistic, and does not consider the contextual structure of the gloss information (e.g. negation words). This results in, for example, the synset arrogance.n.01 being incorrectly labelled as having a positive polarity by their model, due to the positive term 'superior' within its gloss. Upon inspection,

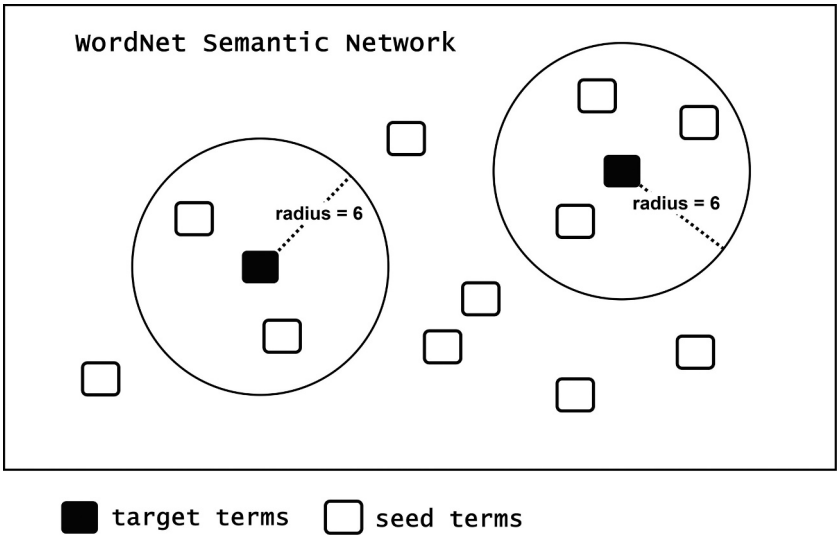

Fig. 1. Coverage issue in semantic relations expansion algorithm.

SentiWordNet 3.0 incorrectly assigns this synset as $Pos(s) = 0.5$, and $Neg(s) = 0.375$.

2.2. Walking WordNet relations

Agerri and García-Serrano (2010) develop Q-WordNet, a lexicon generated using a sense-level polarity assignment algorithm. The algorithm is fully unsupervised, and involves 'walking', or traversing, through WordNet synsets via all semantic relations of all POS categories, assigning a polarity to each synset along its path. A limitation inherent in the Q-WordNet algorithm is that it treats word polarity classification as a 'hard' classification problem, and does not consider polarity strength. Moreover, although it performs well to filter out objective terms, it has a conservative nature in its polarity classification technique, filtering out subtly-subjective terms as objective (i.e. favoring accuracy at the cost of coverage). Additionally, it fails to utilize gloss information during the categorization process. Finally, they use all semantic relations to traverse the graph, but some relations do not effectively preserve sentiment properties, e.g. the *attribute* relation of the synset happiness.n.01 is linked to both happy.a.01 and unhappy.a.01 at the same time.

San Vicente et al. (2014) propose Q-WordNet as Personalized PageRanking Vector (QWN-PPV), employing a WordNet graph and a personalized pagerank algorithm to assign positive or negative labels to synsets. They intrinsically evaluate their final lexicon against the intersecting terms in the General Inquirer lexicon, yielding an average accuracy of 0.75. They do not consider assigning polarity strength scores to senses. Moreover, they use all semantic relations to traverse the graph, but some relations do not effectively preserve sentiment properties.

2.3. Utilizing semantic networks

For the sentiment lexicon induction task, other models have utilized the semantic networks found in hierarchical knowledge bases such as WordNet. Asgari et al. (2019) employ both a sentiment lexicon and a corpus in a vocabulary expansion and unsupervised domain adaptation approach, which can also be applied on almost every language in existence. Darwich et al. (2017) bootstrap via WordNet semantic relations and then use gloss information to label term senses with a polarity. Peng, and Park (2011) propose a fully automatic method to compile a sentiment dictionary using Constrained Symmetric Nonnegative Matrix Factorization, and bootstrap via WordNet's semantic network. Carrillo-de-Albornoz (2012) develop SentiSense, a sense-level affective lexicon which classifies WordNet synsets into 14 different emotion categories.

Williams and Anand (2009) employ a semantic distance-based algorithm, whereby the distance of a word to multiple pos/neg seed pairs in the network define its polarity strength. Hassan, and Radev (2010) propose a random walk model, whereby the hitting time of a word to reach a set of predefined positive and negative seed words in a relatedness graph reflects its polarity/strength. Rao, and Ravichandran (2009) and Blair-Goldensohn et al. (2008) utilize semi-supervised label propagation for the sentiment lexicon generation task; words in the graph adjacent to similar words receive a boost in strength.

2.4. Prominent issues in state-of-the-art models

Several problems inherent in state-of-the-art sense-level lexicon generation models hinder their ability to accurately assign sentiment polarity to term senses. First, they all require manually-labeled seed sets as input, although it can be argued that the primary objective of an automatic lexicon generation model is to avoid human involvement in the first place. Second, the SentiWordNet 3.0 generation model (Baccianella, Esuli, & Sebastiani, 2010) comprises a semantic relations algorithm that only

retrieves senses that are at close proximity to the seed senses, thus overlooking remote senses that are potentially subjective (see Fig. 1). Finally, the model employs a supervised classification step that prioritizes the objective class over other classes, while other models (San Vicente, Agerri, & Rigau, 2014; Agerri, & Garcia-Serrano, 2010) treat objectivity as non-existent entity, and aggressively filter out non-polar senses, resulting in the labeling of subtly subjective senses as objective. This work aims to propose a sense-level sentiment lexicon generation model that attempts to overcome these prominent limitations.

3. Deriving the polarity of term senses using dual-step context-aware in-gloss matching

This section presents all of the steps carried out for construction of the proposed model. The model works at the smallest possible sentiment-carrying text unit, the term sense level, which allows for word-sense disambiguation of polysemous terms. It has been established in the literature that a sentiment lexicon of term senses would be of more benefit compared to a 'coarse' lexicon of terms alone (Akkaya et al. 2009; Wiebe & Mihalcea 2006), since the different senses of a particular term may not possess the same sentiment polarity. For example, salient.a.01[1] is defined as "having a quality that thrusts itself into attention", and carries a positive connotation; while the second sense of the same term (salient.a.02) is defined as "(of angles) pointing outward at an angle of less than 180 degrees", and most would agree is sentiment-neutral or objective.

This also resolves the ambiguity issue that lies in different word classes. For example, the adjective good.a.04 is defined as "deserving of esteem and respect", and is considered positive; while the noun good.n.04 is defined as "articles of commerce", and is considered objective. Another example is mean.a.02, as in "hateful", conveying negativity; and mean.n.01, as in "statistical average", conveying objectivity. Practically, a pre-processing pass of word sense disambiguation can be run on the text to be classified, and the particular term senses in the text can be mapped to their corresponding senses in the sentiment lexicon, for a more fine-grained sentiment classification, as compared to naïve term-level classification.

The creators of WordNet (Miller et al., 1990) mention that working at the term sense level would allow for distinguishing between multiple senses, prior to classification, which can potentially improve accuracy. A sentiment lexicon of term senses would be able to distinguish between "His most *salient* feature is his smile" as positive, and "The *salient* angle of the triangle is small" as objective. A term-level sentiment lexicon, however, would not be able to distinguish between these two different senses. It would instead incorrectly label both senses with whatever polarity is associated with the general term *salient* in its database.

Osgood, Suci, and Tannenbaum (1957) mention that words naturally have a "prior polarity", which is their natural polarity when taken in isolation, independent of any context. The best way to assign a prior polarity to a word is to ask the question "taken out of context, does this word evoke a positive feeling or a negative feeling?" Following this, the 'general' domain is used, and for each term sense, the prior polarity is assumed. The aim here is to develop a general purpose, domain-independent lexicon that labels term senses with their stereotypical polarity when taken out-of-context.

Most term senses are relatively easy to label, as their stereotypical prior polarity is preserved across multiple domains (e.g. *upset, sadly* and *disease* maintain a negative connotation across different domains). However, other terms have a variable polarity that changes across different domains. For instance, the term *predictable* is considered to carry a positive polarity when discussing the stock market, but is considered to carry a negative polarity in a movie or book review. In the case of variable-polarity term senses, these are simply assigned the polarity with which they appear in-context the most often.

Fig. 2. depicts a high level architecture of the model. The input to the algorithm comprises the initial null seed sets ($S^{(pos)}$ and $S^{(neg)}$). The four main algorithms in the figure work successively. The seed set population algorithm (SSPA) uses the sentiment opposition morphology rules (affix patterns) to automatically populate the null seed sets. The semantic relations expansion algorithm (SREA) uses a semantic network to expand the seed sets. The context-aware gloss expansion algorithm (CAGEA) exploits human-defined gloss information to pick up subjective synsets that were overlooked by the previous algorithm. A pos-tagger is used to pos-tag gloss terms, while the in-context negator list is used for handling negation terms within glosses. Finally, the unsupervised sentiment categorization algorithm (USCA) uses the expanded seed sets as training data, in order to categorize all WordNet's synsets as Positive, Negative or Objective.

3.1. Seed set population algorithm

With the aim of minimizing human involvement, the initial positive and negative seed sets are null sets ($S^{(pos)} = \{\}$ and $S^{(neg)} = \{\}$). No known prior work in the literature on sense-level sentiment lexicon generation is initiated with null seed sets. San Vicente, Agerri, and Rigau (2014) and Agerri and García-Serrano (2010) start with a total of six synsets as seeds extracted from the term 'quality', as well as with Turney and Littman's (2003) 14 terms. Baccianella et al. (2010) and Esuli and Sebastiani (2006) manually select a total of 105 synsets (47 positive and 58 negative) as their initial seed sets. Conversely, in this work, rather than using manually defined seed sets as input, the seed sets are generated automatically using a morphologically-inspired approach. This model is considered fully unsupervised, since it is the morphology rules themselves that automatically generate the initial seed terms (training data), and in turn populate the positive seed set $S^{(pos)}$ and the negative seed set $S^{(neg)}$. This is especially valuable in scenarios where manually labeled training data is lacking or scarce.

[1] Note that the convention used to refer to a WordNet synset (term sense) in this work is term.POS.sense-number. In this case, salient.a.01 refers to the first sense of the adjective salient. In terms of POS, adjectives are denoted as 'a', satellite adjectives as 's', adverbs as 'r', nouns as 'n', and verbs as 'v'.

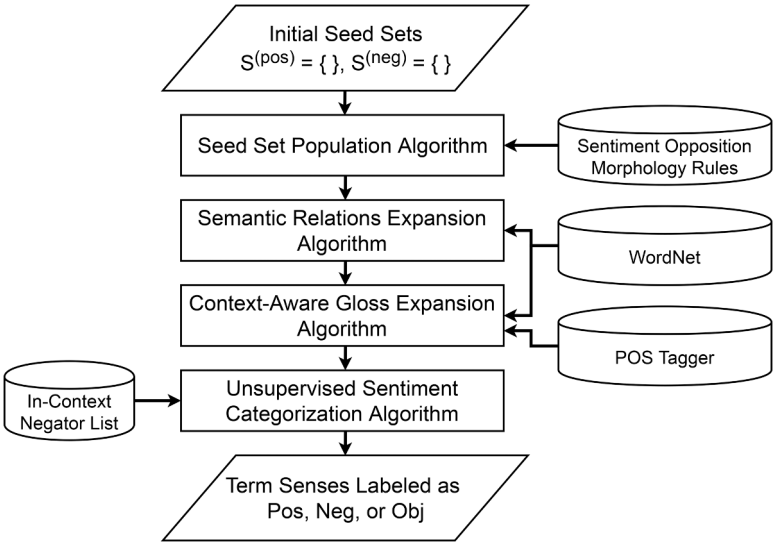

Fig. 2. High-level model architecture.

The subfield of Lexicology in Linguistics deals with the smallest lexical units that carry semantic information (i.e. terms). Generating morphological variations of a particular term involves modifying the morphological structure of the term itself, and in turn manipulating the semantic information encoded within the term. Morphological derivation is the process of forming a new term from an existing term by introducing an affix (Jurafsky & Martin, 2014). This process results in two possible scenarios: (1) the meaning of the original term is preserved and only the syntactic category changes ($happy^{adj}$ vs $happily^{adv}$); or (2) the semantic meaning of the original term is changed and the syntactic category remains the same ($happy^{adj}$ vs $unhappy^{adj}$). Both terms always share the same root, or head term. In this work, the semantic dimension of *sentiment* is of particular interest. This involves the second scenario of morphological derivation mentioned above, namely, modifying the semantic meaning of a term pair, while preserving the syntactic category.

According to the creators of WordNet (Miller et al., 1990): "*most antonyms of adjectives are formed by a morphological rule that inverts the polarity of the meaning by adding a negative prefix (un~, in~, il~, etc.).*" The majority of antonymous word pairs in English are generated by morphological rules (affix patterns) that 'negate' the original term when applied. This stems from the Marking Theory in Linguistics (Battistella, 1990; Lehrer, 1974), which states that the original unmarked term carries a positive connotation, while the marked version of the same term tends to carry a negative connotation. For the majority of antonymous term pairs, the negation marker is explicitly attached to the negative term in the form of a negative affix. The morphological rules that abide by this theory generate term pairs with distinct semantic and lexical properties. Semantically, a generated term pair shares a relation of *sentiment opposition*, and lexically, the term pair shares an *unmarked-marked* relation (e.g. *healthy-unhealthy, honest-dishonest, normal-abnormal*).

Therefore, based on this morphologically inspired method, a set of predefined morphology rules is employed to reliably generate unmarked-marked term pairs, effectively populating $S^{(pos)}$ with unmarked terms, and $S^{(neg)}$ with their marked counterparts. This allows for the generation of two dichotomous seed sets that comprise bipolar term pairs with opposing sentiment polarity.

This has been empirically validated early on in the literature. For example, Hatzivassiloglou and McKeown (1997) test a small subset of automatically extracted unmarked-marked term pairs (*adequate-inadequate*), and conclude that this morphological method is highly accurate, achieving a 97.06% accuracy to automatically assign unmarked terms as positive and marked terms as negative. In this work, this heuristically defined set of morphology rules (affix patterns) is coined as the *Sentiment Opposition Morphology Rules* (SOMRs). The SOMRs only include affix patterns that tend to generate term pairs with opposing polarity, while other affix patterns that are not expected to create an opposing polarity (e.g. *pro~*, *hyper~*, *ultra~*, etc.) were not included.

Table 2 shows the SOMRs, and some example term pairs generated by each rule. The SOMRs were applied on the lexical categories of adjectives and adverbs only, since Miller et al. (1990) observe that binary opposition occurs only when these rules are applied on adjectives and adverbs, and not on nouns and verbs. A test experiment was performed to validate their observation, and the application of these rules on verbs and nouns resulted in many sentiment-neutral terms being added to the seed sets.

The practical operationalization of the seed set population algorithm in pseudocode format is shown in Fig. 3. For every rule in the SOMRs, and for each term sense in WordNet: if the string and its marked counterpart are included in WordNet, and they are both in the same lexical category, the unmarked term sense is added to $S^{(pos)}$, while the marked term sense is added to $S^{(neg)}$, reliably populating the seed sets.

Table 2
SOMRs (affix patterns) used to generate unmarked-marked term pairs.

SOMR	Sample Term Pairs
ab~	(normal, abnormal)
de~	(monetize, demonetize)
dis~	(advantage, disadvantage)
dys~	(function, dysfunction)
il~	(legal, illegal)
im~	(mobilize, immobilize)
in~	(sufficient, insufficient)
ir~	(reverence, irreverence)
mal~	(nourished, malnourished)
mis~	(inform, misinform)
non~	(functional, nonfunctional)
un~	(able, unable)
under~	(develop, underdevelop)

Seed Set Population Algorithm

```
Input:  Null versions of initial S(pos), S(neg)
Output: Populated versions of S(pos), S(neg)
Begin
For each affix in SOMRs:
   For each applicable lexical_category in WordNet:
      For each TERM in lexical_category:
         If (TERM) & (affixTERM) exist in lexical_category
            S(pos) <- S(pos) + TERM
            S(neg) <- S(neg) + affixTERM
         End If
      End For
   End for
End For
S(pos) populated <- S(pos)
S(neg) populated <- S(neg)
End
```

Fig. 3. Seed set population algorithm.

3.2. Semantic relations expansion algorithm

This step involves calling WordNet semantic relations on the populated seed sets (after running the SSPA) to expand them with linked synsets that also have related sentiment properties. This algorithm is based on the intuition that the polarity of a set of paradigm seed terms is preserved as they strategically propagate through the network of predefined semantic relations, since terms linked via these relations not only have related semantic meanings, but also have related sentiment properties. In order to bootstrap the seed sets via the semantic network, new synsets are automatically added to the seed sets based on their semantic relations with other synsets in the network.

To a considerable degree, an assumption can be made that the synonyms of a word have an equal polarity to that word, while its antonyms have an opposing polarity. Synonyms inherit the same polarity, while antonyms inherit the opposite. However, the synonymy relation cannot be used here, since this is the lexical relation that connects synonym member terms to each other within a synset itself. For this reason, the semantic relations among the synsets themselves, and not among the synset member terms within each synset, are utilized.

The semantic relations of (1) see-also, (2) similar-to, (3) direct hyponym, (4) direct troponym, (5) entailment, (6) derivationally related form and (7) pertainym have been empirically observed to preserve polarity during expansion. In this work, these relations are coined as the Sentiment Preserving Semantic Relations (SPSRs). The SPSRs are employed for retrieving new synsets linked to seed set synsets, and then adding these new synsets to the same seed set, since the sentiment polarity is preserved. The aim is to automatically generate continuously expanding 'chains of synsets', while preserving the polarity. For example, happy.a.01 is linked to other positive terms such as cheerful.a.01 and contented.a.01 via the see-also relation, and blessed.a.06 and blissful.a.01 via the similar-to relation.

Conversely, the semantic relation of antonym has been empirically observed to invert polarity during propagation. In this work, this relation is coined as the Sentiment Inverting Semantic Relation (SISR), and is employed for retrieving new synsets linked to seed

```
Semantic Relations Expansion Algorithm
Input:  S(pos), S(neg) output from previous algorithm
Output: Expanded versions of S(pos), S(neg)
Begin
For pos_syn in S(pos):
   For each rel in SPSRs:
     S(pos) <- S(pos) + new_syn linked via rel(pos_syn)
   End For
   For each rel in SISR:
     S(neg) <- S(neg) + new_syn linked via rel(pos_syn)
   End For
End For
For neg_syn in S(neg):
   For each rel in SPSRs:
     S(neg) <- S(neg) + new_syn linked via rel(neg_syn)
   End For
   For each rel in SISR:
     S(pos) <- S(pos) + new_syn linked via rel(neg_syn)
   End For
End For
S(pos) expanded <- S(pos)
S(neg) expanded <- S(neg)
End
```

Fig. 4. Semantic relations expansion algorithm.

synsets, and then adding these new synsets to the opposing seed set, since the polarity is inverted. As an example, happy.a.01 is connected to the negative term unhappy.a.01 via the antonym relation.

During every iteration of this algorithm, the seed sets continue to expand as new synsets in the WordNet network are retrieved and added to them through the mentioned semantic relations. On every iteration, new synsets linked via the SPSRs are added to the same seed set, while new synsets linked via the SISR are added to the opposing seed set.

The farther the distance between seed terms and unseen terms in the network (i.e. the more edges in the path between them), the weaker their semantic relation becomes, and consequently, the higher the volume of 'noise' added to the seed sets. It has been well-established in the literature (Ide, 2006; Wiebe, & Mihalcea, 2006) that the semantic significance of a path decreases as a function of its length from the source synset to the destination synset in the WordNet semantic network. It is thus critical to determine a suitable trade-off point between coverage and correctness, in order to expand the seed sets as much as possible, but without compromising accuracy. Baccianella et al. (2010) use a maximum of six iterations of semantic relations expansion in their work in the generation of SentiWordNet 3.0. We also employ a maximum of six iterations of semantic relations expansion, in order to maintain a reasonably level playing field.

Fig. 4. presents the semantic relations expansion algorithm in pseudocode format. During every algorithm iteration, for each positive synset (pos_syn) in $S^{(pos)}$: all of the synsets connected to pos_syn via the SPSRs are added to $S^{(pos)}$, while all of the synsets connected to pos_syn via the SISR are added to $S^{(neg)}$. During the same iteration, simultaneously, the exact same analogous process is performed for each negative synset (neg_syn) in $S^{(neg)}$.

3.3. Context-aware gloss expansion algorithm

As mentioned, an inevitable problem inherent in the previous semantic relations expansion algorithm is that it is only able to retrieve terms that are at a close proximity to the seed terms in the network. A maximum of six iterations was chosen for semantic relations expansion in the previous step, which was heuristically observed to be the optimal number of iterations to increase coverage, with as minimum effect as possible on accuracy. In order to provide a solution to this issue, an additional round of seed set expansion is performed by 'mining' the human-defined gloss information in a digital dictionary. This is deemed as the context-aware gloss expansion algorithm (CAGEA), and ensures synsets that were beyond the reach of the semantic relations expansion algorithm are identified and added to the seed sets.

It is vital to note that by 'context', we refer to the textual context of the gloss information in which the target sense to be classified appears in. Therefore, the model is 'context-aware' in that it utilizes the context of the glosses a sense appears in, in order to decide whether that particular sense carries a positive connotation or a negative one. For instance, the target sense happy.a.04 appearing in the gloss of glad.s.03, a sense in the positive seed set, is considered to have appeared in a 'positive context', hence is considered to lean towards positivity.

In this algorithm, the definiendum-definiens semantic relation (i.e. the relation between an entry itself (definiendum) and a constituent term within its gloss (definiens)) is exploited with the intuition that a constituent term contained within the gloss of an entry is related to the entry term semantically. Therefore, terms within the gloss of the entry are deemed to generally have a similar semantic orientation to that of the entry itself. For instance, the gloss of worthy.a.01 is formally defined as "having worth or merit or value; being honorable or admirable", which contains many positive terms such as merit, value, honorable and admirable. A POS

In-Context Negator List
not, no, none, nothing, neither, nowhere, never, cannot, free from, without, nobody, no one, fail, failure, violating, deficient, devoid, lack, lacks, lacking, unlikely, hardly, scarcely, barely

Fig. 5. In-context negator list.

tagging step on the gloss terms is performed, prior to the gloss matching operation, in order to allow a match if both terms are under the same lexical category.

Some synsets include negation terms within their glosses. Negation is defined as a grammatical category that has the ability to modify the truth value of a target proposition (Zhu, Guo, Mohammad, & Kiritchenko, 2014; Morante, & Sporleder, 2012; Councill, McDonald, & Velikovich, 2010; Wiegand, Balahur, Roth, Klakow, & Montoyo, 2010), and is often exhibited by the use of explicit negation terms (not, never, cannot, etc.). Therefore, in this algorithm, a negation rule is applied in order to avoid adding a synset to the wrong seed set by overlooking negated terms. For example, without a negatitableTon rule, the synset undiplomatic.a.01 is defined as "not skilled in dealing with others", and would be incorrectly added to the positive seed set because of the positive term skilled. To counteract this negation issue, a negation rule is applied to detect that the term skilled is under the contextual scope of the negation term not, and would thus not be added to the positive seed set. The in-context negator list includes explicit negation terms that are used by the algorithm (no, not, nothing, cannot, etc.), as well as implicit terms and phrases which refer to negation in a general sense (lack of, free from, scarcely, etc.). Fig. 5. presents the in-context negator list used by this algorithm.

Context-Aware Gloss Expansion Algorithm

```
Input:  S(pos), S(neg) output from previous algorithm
Output: Expanded versions of S(pos), S(neg)
Begin
For syn in WordNet and not in S(pos), S(neg):
   //Step 1: In-gloss matches within candidate syn glosses
   For each adj_syn of all adj_syns in gloss[syn]:
      For pos_syn of all pos_syns:
         If match(adj_syn & pos_syn) & match NOT preceded by negator
            pos_match_count ++
         End If
      End For
      For neg_syn of all neg_syns:
         If match(adj_syn & neg_syn) & match NOT preceded by negator
            neg_match_count ++
         End If
      End For
   End For
   //Step2: In-gloss matches within seed syn glosses
   For gloss[pos_syn] of all pos_syns:
      For each adj_syn in gloss[pos_syn]:
         If match(syn & adj_syn) & match NOT preceded by negator
            pos_match_count ++
         End If
      End For
   End For
   For gloss[neg_syn] of all neg_syns:
      For each adj_syn in gloss[neg_syn]:
         If match(syn and adj_syn) & match NOT preceded by negator
            neg_match_count ++
         End If
      End For
   End For
   If pos_match_count > neg_match_count
      S(pos) <- syn
   Else If neg_match_count > pos_match_count
      S(neg) <- syn
   Else
      discard syn
   End If
End For
End
```

Fig. 6. Context-aware gloss expansion algorithm.

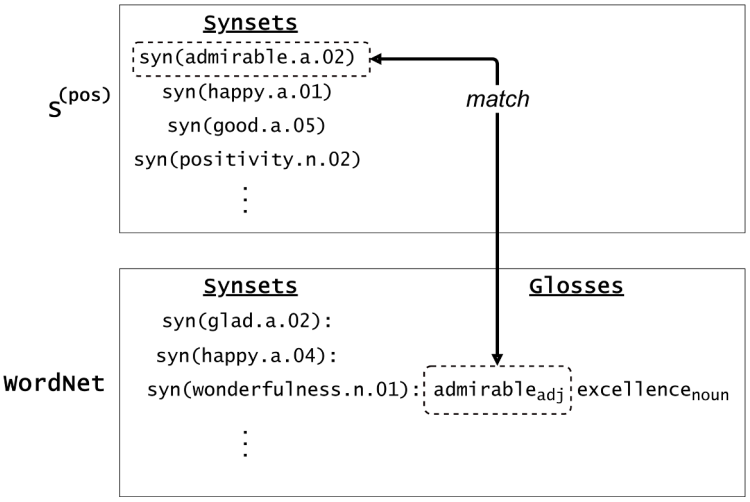

Fig. 7. In-gloss matches within candidate syn glosses.

Fig. 6. shows the full version of the CAGEA algorithm in pseudocode format. The input to this algorithm is the set of WordNet synsets that are not in $S^{(pos)}$ or in $S^{(neg)}$. The output from this algorithm is the expanded versions of $S^{(pos)}$ and $S^{(neg)}$.

Each WordNet synset (syn) not in $S^{(pos)}$or in $S^{(neg)}$ is input into the algorithm. This algorithm involves a dual-step in-gloss matching process. Step 1 involves in-gloss matches within candidate syn glosses, while Step 2 involves in-gloss matches within pos_syn and neg_syn glosses. Note that the latter is an inverse of the former. During the first in-gloss matching step, all of the adjective synsets (adj_syns) within the gloss of the target syn are extracted, under the condition they are not preceded by a negation term. Next, each of these adj_syns is compared to each pos_syn in $S^{(pos)}$ and to each neg_syn in $S^{(neg)}$, in order to check for possible in-gloss matches. The frequency of all adj_syns in $S^{(pos)}$ are recorded in pos_match_count, and the frequency of all adj_syns in $S^{(neg)}$ are recorded in neg_match_count. Fig. 7. illustrates an example of an in-gloss match between an adj_syn within the gloss of syn, and a pos_syn in $S^{(pos)}$. In this case, the pos_match_count for the syn wonderfulness.n.01 is incremented by 1, due to the match between the adj_syn admirable within its gloss, and the pos_syn admirable.a.02 in $S^{(pos)}$.

The reason only adjectives are extracted from synset glosses in this step is because, in linguistics, this POS class contains the largest amount of subjective terms. According to Esuli and Sebastiani (2006), 35.7% of adjectives express sentiment, while only 9.98% of nouns and 11.04% of verbs express sentiment. In line with their claim, using gloss synset terms with a POS other than adjectives resulted in the addition of noisy data to the seed sets, due to in-gloss matches that did not necessarily preserve the sentiment properties of the newly added synsets.

Simultaneously, during the same run, the second in-gloss matching step is performed. First, the entry synset itself (syn) is compared to the adj_syns in the gloss of each pos_syn in $S^{(pos)}$, and also to the adj_syns in the gloss of each neg_syn in $S^{(neg)}$. Fig. 8. illustrates an example of an in-gloss match between a syn and an adj_syn in the gloss of a pos_syn. For each match between syn and an

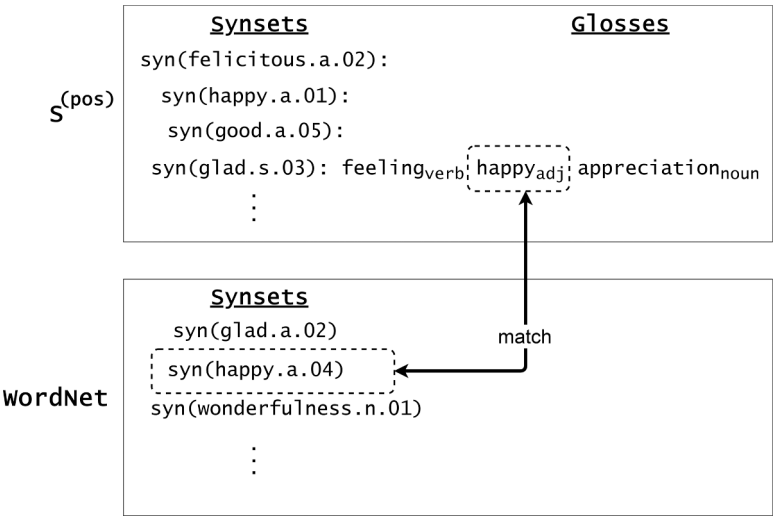

Fig. 8. In-gloss matches within seed pos_syn and neg_syn glosses.

adj_syn within each pos_syn, pos_match_count is incremented; and for each match between syn and an adj_syn within each neg_syn, neg_match_count is incremented. Note that this second in-gloss matching step allows for the retrieval of only adjectives to seed sets.

After both in-gloss matching steps have been performed, pos_match_count and neg_match_count are compared. If pos_match_count is greater than neg_match_count, syn is added to $S^{(pos)}$; else if neg_match_count is greater than pos_match_count, syn is added to $S^{(neg)}$. Otherwise, if the match count of syn is equal in both classes, or it is not matched in either class, then it is considered sentiment-neutral, and discarded. This algorithm is run iteratively until there are no more synsets to pick up, and the seed sets $S^{(pos)}$ and $S^{(neg)}$ are fully expanded.

3.4. Unsupervised sentiment categorization algorithm

All of the remaining synsets that are not included in the fully expanded $S^{(pos)}$ and $S^{(neg)}$, are considered to be objective, and are added to the objective class $S^{(obj)}$. For simplicity, the objective class is heuristically defined as $S^{(obj)} = \text{WordNet} - (S^{(pos)} \cup S^{(neg)})$. After all three seed sets are defined, they are used as the final training data.

The unsupervised sentiment categorization algorithm (USCA) labels all WordNet synsets as Positive, Negative or Objective, along with a numerical score representative of a measure of confidence of the categorization decision. Multiple occurrences in the seed sets are retained, and a log of the synset frequency counts is recorded for each synset. Centrality measures in the bounds of [1, -1] act as measures of confidence a synset belongs to its respective category.

According to the Social Network Theory (Burt, 1980), a node that has many semantic relations with other nodes that are members of a particular category, also tends to be more central to that category. Moreover, in the related literature, semantic connectivity between terms in a semantic network has been proven to be an important indicator of orientation. Hassan and Radev (2010) practically demonstrate that connectivity via semantic relations is more effective compared to other indicators such as semantic distance, or the length of the path between two terms in the network (Williams, & Anand, 2009).

Therefore, based on the Social Network Theory, and on the concept of term connectivity via semantic relations, the confidence measure (conf) for all WordNet synsets is computed. For each synset, all of its neighboring synsets are extracted, and the conf of the synset is computed based on the occurrence frequency of it and its 'neighbors' in $S^{(pos)}$ and $S^{(neg)}$. The 'neighbor' of a target synset is defined as any synset linked to the target synset via the semantic relations of *similar-to, also-see* and *derivationally-related-form*, since these relations have been observed to optimally preserve sentiment properties during propagation in the semantic network, compared to other relations. Moreover, the mentioned relations apply for all POSs. In other words, any synset one hop away from the target synset via these relations is considered to be a neighbor of the target synset, and can be used to aid in the computation of the conf of the target synset. After a practical investigation, it was observed that utilizing the neighbors of a target synset in the computation of conf has yielded better results compared to using the occurrence frequency of the synset alone. Using $S^{(pos)}$ and $S^{(neg)}$ as training data, the conf of every WordNet synset (syn) is computed as follows:

$$conf_{pos} = freq(syn^{(pos)}) + \sum_{n \in neighbors(syn)} freq(n^{(pos)}) \tag{4.1}$$

$$conf_{neg} = freq(syn^{(neg)}) + \sum_{n \in neighbors(syn)} freq(n^{(neg)}) \tag{4.2}$$

$$conf_{final} = \frac{conf_{pos} - conf_{neg}}{conf_{pos} + conf_{neg}} \tag{4.3}$$

where $freq(syn^{(pos)})$ represents the frequency of syn in $S^{(pos)}$, and $\sum_{n \in neighbors(syn)} freq(n^{(pos)})$ represents the summation of all the frequencies of each neighbor (n) in $S^{(pos)}$. The same applies for$S^{(neg)}$. Note that $neighbors(syn)$ is representative of all synsets that have direct links to syn via the mentioned relations. The denominator represents the total frequency of syn and its neighbors in both classes, and acts as a normalization mechanism to constrain the final conf score to the bounds of [1, -1]. Finally, if the conf of a syn is positive, then it is labeled as Positive in the sentiment lexicon, along with its numerical conf score. The same applies for Negative. If it is a zero-value, then it is sentiment-neutral, and labeled as Objective.

4. Experimental setup and evaluation

The experimental setup involves both intrinsic and extrinsic evaluation of the proposed model. The former involves computing the model's overall accuracy, and then comparing it to that of related models on the same gold standard benchmarks. The latter involves practically employing the generated sentiment lexicon in a 'real-world' polarity classification task on four publically-available datasets of varying domains, and comparing its performance to that of other state-of-the-art sense-level sentiment lexicons, as well as to a naïve term-level version of the same lexicon.

4.1. Intrinsic evaluation

Two independent intrinsic evaluation procedures were employed. The first computes the model's accuracy against the General

Inquirer lexicon[2] (GI; Stone, Dunphy, & Smith, 1966), while the second computes the model's accuracy against the MicroWN(OP)-3 lexicon[3] (MWN; Baccianella et al., 2010).

The GI was chosen to allow a feasible comparison to state-of-the-art models, namely, the QWN-PPV lexicon generation model San Vicente et al., 2014, the SentiWordNet 1.0 generation model (Esuli, & Sebastiani, 2006), and the SentProp lexicon generation model (Hamilton et al. 2016).

Due to its widespread reliability, it has been widely applied as a gold standard benchmark for evaluation purposes in the majority of prior works on term-level and sense-level sentiment lexicon generation. The GI is a manually constructed lexicon that comprises terms categorized based on their semantic properties. A term may be included in multiple categories, each one denoting a particular trait associated with the term. Among the available categories are 'Positiv' and 'Negativ' (sic), which consist of 1915 positive terms and 2291 negative terms respectively. The remaining 7582 terms not labeled with a polarity can be implicitly deemed as 'Objective'. After removing multiple occurrence of the same term due to multiple senses, these sets were reduced to 1637 positive, 2007 negative, and 5344 objective terms.

Note that although the GI is a term-level lexicon, San Vicente et al. (2014) and Esuli and Sebastiani (2006) use it to evaluate their sense-level lexicon generation models. Therefore, this evaluation also follows exactly the same evaluation procedure they employed. In the generated lexicon, all term senses of the same POS are collapsed to a single occurrence in the lexicon, and the polarity most prominent among all senses is assigned to the collapsed term. The set of terms that intersect between the generated lexicon and the GI are extracted, and the polarity labels in the generated lexicon are compared to those in the GI, in order to compute the accuracy of the model.

The second manually annotated benchmark, the MWN lexicon, was used in order to allow for a feasible comparison to the SentiWordNet 3.0 generation model. This benchmark comprises 1105 WordNet 3.0 synsets that were carefully manually-labelled by Baccianella et al. (2010) for *Pos(s), Neg(s)* and *Obj(s)* scores. The set of term senses that intersect between the generated lexicon and MWN were extracted, and the polarity labels in the generated lexicon were compared to those in MWN, in order to compute the overall accuracy of the model. Since each term sense in MWN is assigned both a positive score and a negative score simultaneously, the final polarity of a term sense taken is simply the polarity label with the higher value. All synsets with no positive and negative values, or equal positive and negative values, were considered objective. After assigning these 1105 synsets a final positive, negative or objective polarity, the final test set used comprises 276 positive, 267 negative, and 512 objective synsets.

4.2. Extrinsic evaluation

Although intrinsic evaluation allows for measuring the accuracy of a sentiment lexicon induction model, this is not a necessarily a suitable indication that the resultant lexicon would perform well in sentiment analysis tasks on large pieces of text (San Vicente et al., 2014; Mohammad, Dunne, and Dorr, 2009). After all, a sentiment lexicon is as beneficial as its results in a practical sentiment analysis task.

The four datasets used for the classification task are as follows. The first is the Stanford Large Movie Reviews Dataset (SLMRs; Maas, Daly, Pham, Huang, Ng, & Potts, 2011), focusing on the movie reviews domain. From this dataset, we have used 5k positive and 5k negative manually labelled movie review documents. The second dataset is the Book Reviews Multi-Domain Sentiment Dataset (BRMSD; (Blitzer, Dredze, & Pereira, 2007), focusing on the book reviews domain. This dataset contains 1k positive and 1k negative manually labelled book review documents. The third dataset is the Electronics Reviews Multi-Domain Sentiment Dataset (ERMSD; Blitzer et al., 2007), focusing on the electronic reviews domain. This dataset contains 1k positive and 1k negative manually labelled electronics review documents. The fourth dataset is the Sports Equipment Reviews Multi-Domain Sentiment Dataset (SERMSD; Blitzer et al., 2007), focusing on the sports equipment reviews domain. This dataset contains 1k positive and 1k negative sports equipment review documents.

Extrinsic evaluation involves measuring the standard evaluation metrics of precision (P), recall (R) and F-measure (F) achieved by the generated lexicon in a sentiment classification task. P is formally defined as the portion of retrieved items that are relevant; R is the fraction of the relevant items that have been successfully retrieved; and F represents the harmonic mean of both precision and recall. In this work, $P = TP/(TP + FP)$, $R = TP/(TP + FN)$, and $F = 2PR/PR$. TP represents the items correctly classified as positive (true positives), *FP* the items incorrectly classified as positive (false positives), and *FN* the items incorrectly classified as negative (false negatives).

The first evaluation procedure aims to compare the generated lexicon to prominent state-of-the-art sense-level lexicons, namely, QWN-PPV (San Vicente et al., 2014), QWordNet (Agerri, & Garcia-Serrano, 2010) and SentiWordNet 3.0 (Baccianella et al., 2010)(. In these works, each lexicon was generated and structured in a slightly different layout from the proposed lexicon. The structure was modified in order to better facilitate a side-by-side comparison to the proposed lexicon, but without modifying the content of the lexicons used for comparison.

Regarding SentiWordNet, if the positive score of a synset is higher than the negative score, it is considered as having a positive polarity, and the opposite is true for synsets with a negative polarity. Objective synsets are discarded. This amounted to 13,128 positive and 14,726 negative synsets in the final lexicon used. Regarding QWordNet, the lexicon generated using WordNet 3.0 was used, containing 4402 positive and 8108 negative synsets. Regarding QWN-PPV, the lexicon achieving highest accuracy in their

[2] http://www.wjh.harvard.edu/~inquirer/

[3] https://github.com/aesuli/SentiWordNet/blob/master/data/Micro-WNop-WN3.txt

evaluation was used, namely, QWN-PPV-TL (s05_G4), which means that the 5th iteration of synset seeds was used to propagate via the G4 graph (i.e., every semantic relation except antonomy and glosses). This final lexicon contains 56,343 positive and 58,355 negative synsets.

In this simplified polarity classification task on the four datasets considered, a positive match is considered if a positive sense in the lexicon matches a sense in the document, and is assigned a +1 value. Similarly, a negative match is considered if any negative sense in the lexicon matches a sense in the document, and is assigned a -1 value. For every document in the corpus, the final polarity is simply computed as the number of positive matches found in the text, minus the number of negative matches, as follows:

$$polarity(doc) = \sum_{sense \in doc} polarity(sense)$$

where *polarity*(*sense*) is the polarity of each term sense (*sense*) in the document *doc*.

Indeed, adding strategic syntactic rules to the classification model would certainly generate more impressive results. However, a simple explicit matching technique is used to prohibit the effect other factors have on classification. Only a preprocessing pass of word sense disambiguation[4] is performed on each document before categorization, in order to map each term in the corpus to its corresponding term sense (synset) in WordNet. Finally, the P, R and F values achieved by the generated sense-level lexicon are compared to the values achieved by the state-of-the-art sense-level lexicons available in the literature.

The second evaluation procedure aims to practically demonstrate the advantage a finer-grained sense-level lexicon has over a standard term-level lexicon, in a sentiment polarity classification task on the same datasets employed above. In the first step, only term matching is performed between a term-level version of the generated lexicon, and the test set of manually labeled positive and negative documents. In this case, all term senses of the same pos are collapsed to a single occurrence in the lexicon, and the polarity most prominent among all senses is assigned to the collapsed term. In the second step, sense-matching is performed between the original generated sense-level lexicon, and the sense-disambiguated version of the dataset. In this case, a preprocessing step of word sense disambiguation (WSD) on the test set is performed to map each document term with its corresponding sense in WordNet, allowing for a sense-disambiguated version of the dataset.

In this simplified polarity classification task. The same equation above is used to compute the overall polarity of each review document in the same manner as in the first procedure. Finally, the P, R and F values achieved by the generated sense-level lexicon are compared to the values achieved by the term-level version of the same lexicon.

5. Results and discussion

This section presents the results from both intrinsic and extrinsic evaluation of the proposed model, as well as a comprehensive discussion based on these results. Section 5.1 presents the results from intrinsic evaluation, while Section 5.2 presents the results from extrinsic evaluation.

5.1. Intrinsic evaluation results

5.1.1. Results against general inquirer gold standard

Table 3 presents the accuracy of the model against the General Inquirer (GI), for ternary positive-negative-objective categorization, while Fig. 9. shows these plots on a graph. The top three performing SOMRs are reported, namely, lex^i_{dis}, lex^i_{un} and lex^i_{in}, with an average accuracy of 0.721, 0.651 and 0.634, and the best performing accuracy of 0.754, 0.708 and 0.685 respectively. It is critical to note that these three top performing SOMRs are ranked for accuracy based on the best performing model parameters (i.e. SOMR applied, and number of iterations), which is in line with SentiWordNet (Esuli & Sebastiani, 2006).

Several important points are highlighted. First, according to the results, the model's accuracy to categorize negative synsets is always higher compared to its accuracy to categorize positive synsets. This is likely because the English language contains significantly more positive synsets. Second, the accuracy of the model to categorize subjective (positive and negative) synsets is higher than its accuracy to categorize objective synsets. For lex^i_{dis}, the average accuracy of subjective synsets vs objective synsets is 0.731 vs 0.696; for lex^i_{in}, it is 0.635 vs 0.564; and for lex^i_{un}, it is 0.670 vs 0.584.

Third, the best performing accuracy for all SOMRs is after only one iteration of semantic relations expansion, primarily due to the addition of noisy data with an increase in any subsequent iterations. Across all lexicons generated, as the number of iterations increases, accuracy decreases, since the semantic relations in the network become weaker as the distance from the source synset to the destination synset is increased. The only exception is lex^i_{in} decreases in coverage after the 5th iteration, while lex^i_{un} decreases in coverage after the 3rd iteration. This is due to the fact that the training data becomes higher than the pool of subjective synsets to pick up in WordNet. Notwithstanding this exception, the observation holds that when the training data is relatively large, coverage always increases at the cost of accuracy. This observation is in line with several related works (San Vicente et al., 2014; Agerri, & García-Serrano, 2010; Baccianella et al., 2010; Hassan, & Radev, 2010). Using the lex^1_{dis} generation model, there is a significant increase in coverage, at the cost of a slight decrease in accuracy. This is shown in the coverage vs accuracy plots for lex^1_{dis} in Fig. 10.

The only available sense-level lexicon generation models for comparison on this benchmark are the QWN-PPV lexicon generation

[4] The pywsd module for Python was used for mapping terms to their corresponding WordNet synsets. It can be accessed at https://github.com/alvations/pywsd.

Table 3
Accuracy against GI for ternary sentiment categorization.

Lex	lex^i_{dis}				lex^i_{in}				$lexn^i_u$			
	Pos	neg	obj	all	pos	neg	obj	all	pos	neg	obj	all
lex^1_k	.717	.798	.749	.754	.607	.73	.719	.685	.587	.796	.742	.708
lex^2_k	.691	.772	.738	.734	.581	.735	.729	.682	.575	.819	.728	.707
lex^3_k	.698	.798	.703	.730	.561	.730	.712	.668	.548	.818	.664	.677
lex^4_k	.623	.814	.701	.712	.525	.752	.61	.629	.532	.85	.522	.635
lex^5_k	.620	.817	.679	.705	.491	.746	.501	.579	.488	.818	.485	.597
lex^6_k	.614	.811	.665	.696	.453	.715	.526	.564	.436	.776	.539	.584

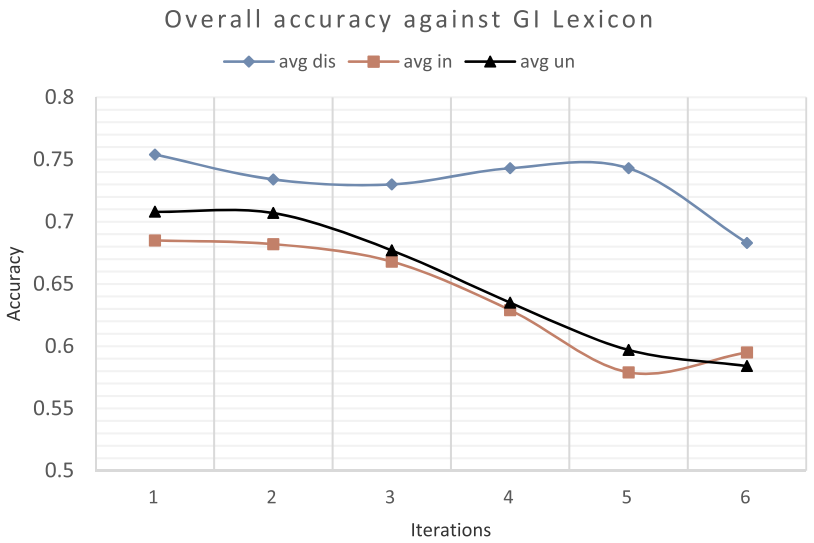

Fig. 9. Accuracy against GI plots for ternary sentiment categorization.

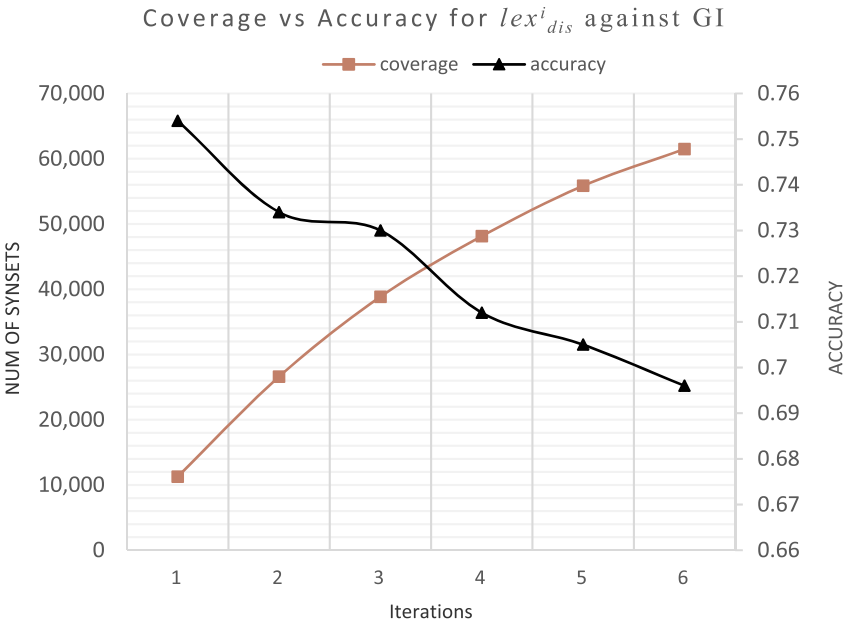

Fig. 10. Coverage vs accuracy against GI plots for lex^i_{dis}.

Table 4
Comparison to prior work on the GI lexicon.

Model	Supervision	Classes	Accuracy
lex^{1}_{dis}	unsupervised	ternary	**0.754**
QWN-PPV	unsupervised	binary	0.750
SentiWordNet 1	supervised	ternary	0.660
SentProp	label propagation	ternary	0.718

model (San Vicente et al., 2014), the SentiWordNet 1.0 lexicon generation model (Esuli and Sebastiani 2006), and the SentProp lexicon generation model (Hamilton et al., 2016). Table 4 shows a comparison between the accuracy of the proposed model vs these models on this same benchmark.

The best-performing lex^{1}_{dis} generation model is comparable to that by San Vicente et al. (2014) (0.754 vs 0.750). It is important to note that their model was evaluated for binary (positive-negative) classification only, whereas the proposed model was evaluated for ternary (positive-negative-objective) classification, practically proven to be a relatively harder task. For example, prior work demonstrated a drop in accuracy when moving from binary to ternary classification by including the objective class. Esuli and Sebastiani (2006) witness a 17% decline in accuracy with the inclusion of the objective class.

The lex^{1}_{dis} generation model performs with an accuracy 9.4% higher than that by Esuli and Sebastiani (2006) (0.754 vs 0.660), although their model requires a committee of eight supervised classifiers (SVMs, Naïve Bayes and Roccio classification models), whereby each individual classifier performs a two-step classification task. In contrast, the proposed model is fully unsupervised and does not require any time-consuming supervised classifiers. Moreover, their model initiates with a total of 105 manually-labeled synsets as the initial (seed sets) training data, while the proposed model initiates with null seed sets. They apply the semantic relations algorithm in their model, which, as pointed out earlier, does not pick up remote synsets, since the semantic relations become weaker as the path between the seed synset and the target synset increases. It is crucial to highlight that the proposed context-aware gloss expansion algorithm has successfully addressed this limitation, whereby in-gloss matching was able to pick up remote potentially subjective synsets that have been overlooked by the semantic relations algorithm. For example, in the SentiWordNet 1.0 lexicon, the synset evil-looking.a.01 is considered objective, since it was overlooked by the semantic relations expansion, but was assigned with a positive orientation by the context-aware gloss expansion algorithm in the proposed model.

The lex^{1}_{dis} generation model performs with an accuracy 3.6% higher than the accuracy obtained by the SentProp lexicon generation model by Hamilton et al. (2016) (0.754 vs 0.718). It is important to note that their model operates at the term level, while ours operates at the sense-level, and that they use a corpus-based approach, while we rely on a (relatively much less data-intensive) digital dictionary as the only source of input. This particular SentProp variant was recreated using the publically-available SocialSent package[5], using off-the-shelf Google news embeddings constructed from 10^{11} tokens. They refer to this as the 'Standard English' variant in their work. We refrain from using other SentProp variants such as those generated using embeddings from financial documents, or from Twitter-specific corpora, since they are domain-specific. This result demonstrates that the proposed model is able to also outperform models that rely on text corpora.

Adreevskaia and Bergler (2006) empirically measure the intersection of agreement between two independent human annotators on the same test set, and conclude that the accuracy of human agreement on the task of labeling words is 0.787. The accuracy achieved by the proposed model is only 3.3% inferior to human performance in the labeling of polarity to words (0.754 vs 0.787). After evaluation of the model on the GI lexicon, reaching a value near this human baseline is sufficient to validate the performance of the proposed lexicon generation model.

5.1.2. Results against Micro-WN(Op)-3 Gold Standard

Table 5 presents the accuracy of the model against the Micro-WN(Op)-3 (MWN) gold standard benchmark, for ternary positive-negative-objective categorization, while Fig. 11. shows these plots on a graph. The top performing is lex^{i}_{dis}, followed by lex^{i}_{un} and lex^{i}_{in}, with an average accuracy of 0.729, 0.661 and 0.660, and the best performing accuracy of 0.765, 0.724 and 0.728 respectively.

Several important points are highlighted. First, in terms of lex^{i}_{un} and lex^{i}_{in}, the accuracy to categorize negative synsets is always higher compared to its accuracy to categorize positive synsets, while the opposite is true for lex^{1}_{dis}. Second, the accuracy of the model to categorize subjective (positive and negative) synsets is higher than its accuracy to categorize objective synsets. For lex^{1}_{dis}, the average accuracy of subjective synsets vs objective synsets is 0.777 vs 0.634; forlex^{i}_{in}, it is 0.738 vs 0.505; and forlex^{i}_{un}, it is 0.749 vs 0.485.

Third, the best performing accuracy for all SOMRs is after only one iteration of semantic relations expansion, primarily due to the addition of noisy data with an increase in any subsequent iterations. A previously mentioned, across all lexicons generated, as the number of iterations increases, the coverage increases, but at the cost of a decrease in accuracy. The only exception is that lex^{i}_{in} decreases in coverage after the 5th iteration, while lex^{i}_{un} decreases in coverage after the 3rd iteration. This is due to the fact that the training data becomes higher than the pool of subjective synsets to pick up in WordNet. Notwithstanding this exception, the observation holds that when the training data is relatively large, coverage always increases at the cost of accuracy. Using the best-performing lex^{1}_{dis} generation model, there is a significant increase in coverage, at the cost of a slight decrease in accuracy. This is

[5] https://nlp.stanford.edu/projects/socialsent/

Table 5
Accuracy against MWN for ternary sentiment categorization.

Lex	lex^i_{dis}				lex^i_{in}				$lexn^i_u$			
	pos	neg	obj	all	pos	neg	obj	all	pos	neg	obj	all
lex^1_k	.820	.753	.722	.765	.750	.820	.613	.728	.730	.820	.624	.724
lex^2_k	.800	.760	.697	.752	.720	.800	.608	.709	.720	.860	.593	.724
lex^3_k	.820	.750	.654	.741	.690	.820	.579	.696	.690	.820	.528	.679
lex^4_k	.820	.760	.642	.740	.670	.800	.450	.640	.660	.810	.412	.627
lex^5_k	.790	.780	.520	.696	.630	.790	.390	.603	.610	.860	.358	.609
lex^6_k	.720	.760	.571	.683	.600	.770	.392	.587	.570	.844	.400	.604

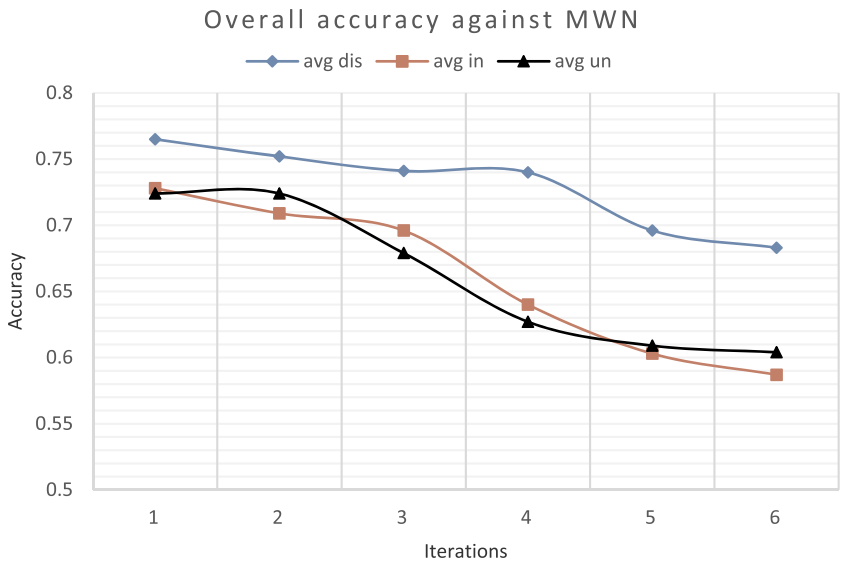

Fig. 11. Accuracy against MWN plots for ternary sentiment categorization.

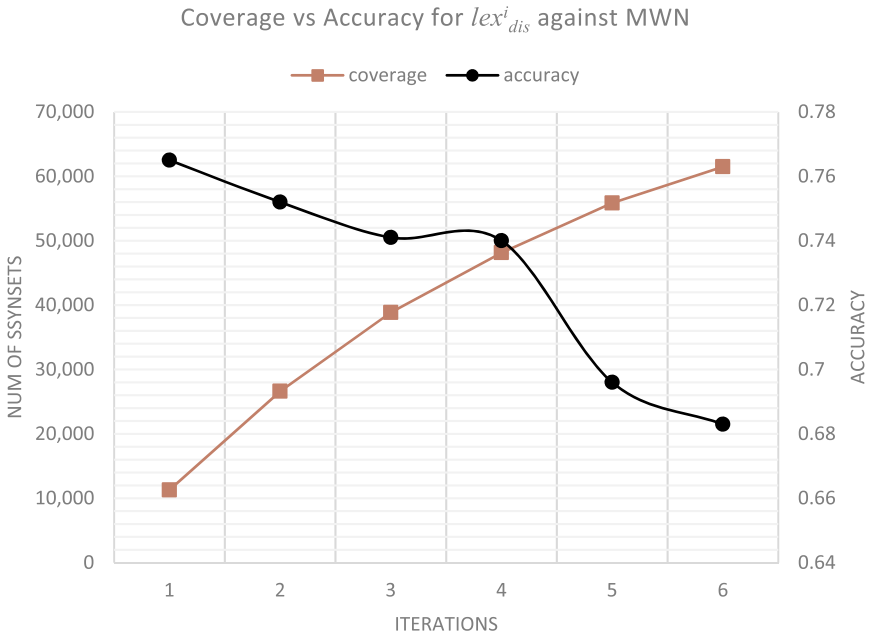

Fig. 12. Coverage vs accuracy against MWN plots for lex^1_{dis}.

Table 6
Comparison to prior work on the MWN lexicon.

Model	Supervision	Classes	Accuracy
lex^{1}_{dis} generation model	unsupervised	ternary	**0.765**
SentiWordNet 3.0	supervised	ternary	0.744

shown in the coverage vs accuracy plots for lex^{1}_{dis} in Fig. 12.

The only available sense-level lexicon generation models for comparison on this benchmark is the SentiWordNet 3.0 generation model (Baccianella et al., 2010)(). Table 6 shows a comparison between the accuracy of the proposed model vs the SentiWordNet 3.0 generation model.

The best-performing lex^{1}_{dis} generation model outperforms the SentiWordNet 3.0 generation model by a very small margin (0.765 vs. 0.744). Notwithstanding this small margin, it is worthy to highlight that their model requires a committee of eight supervised classifiers (SVMs and Roccio classification models), whereby each individual classifier performs a two-step classification task. In sharp contrast, the proposed model is fully unsupervised and does not require any time-consuming supervised classifiers nor the hand-annotated training data associated with them. Their model also employs two time-consuming random walk steps based on an in-gloss matching method which does not consider the contextual structure of the gloss information (e.g. negation words), whereas the proposed model is context-aware.

Moreover, their model initiates with a total of 105 manually-labelled synsets, while the proposed model initiates with null seed sets. The only input into the model is one SOMR (e.g. dis∼), and the model runs in a fully automated manner, from the initiation to its termination, without the reliance on any other human input or involvement.

Furthermore, applying the resultant generated lexicons side by side in this polarity classification task demonstrates that the proposed context-aware gloss expansion algorithm seems to have potentially addressed the limitation inherent in the semantic relations expansion algorithm by picking up remote polar senses in the network. They apply the semantic relations algorithm in their model, which, as pointed out earlier, fails to detect remote polar synsets beyond its reach, since the semantic relations become weaker as the path between the seed synset and the target synset increases. It is crucial to highlight that the proposed context-aware gloss expansion algorithm has successfully addressed this limitation, whereby in-gloss matching is able to pick up remote potentially subjective synsets that are overlooked by the semantic relations algorithm in the SentiWordNet 3.0 generation model. This is because the proposed context-aware gloss expansion algorithm relies on semantic information within glosses, and is not constrained to semantic distance within the semantic network.

The performance of the proposed model against the MWN is on par with human annotation, yielding an accuracy that is only 2.2 % inferior to human performance in the labeling of polarity to words (0.765 vs 0.787). After evaluation of the model on the MWN lexicon, reaching a value near this human baseline is sufficient to validate the performance of the proposed lexicon generation model.

SYN ID	SYN POS	SYN	SYN TERMS	CONF	GLOSS
64479	a	Synset('advantageous.a.01	['advantageous']	1	giving an advantage
64787	s	Synset('beneficial.s.01')	['beneficial', 'good']	1	promoting or enhancing well-being
65064	s	Synset('plus.s.02')	['plus', 'positive']	1	involving advantage or good
65184	s	Synset('discriminatory.s.04	['discriminatory', 'preferential']	1	manifesting partiality
67038	a	Synset('advisable.a.01')	['advisable']	1	worthy of being recommended or suggested; prudent or wise
67379	s	Synset('better.s.03')	['better', 'best']	1	(comparative and superlative of `well') wiser or more advantageous and hence advisable
67638	s	Synset('well.s.03')	['well']	1	wise or advantageous and hence advisable
69531	a	Synset('aesthetic.a.02')	['aesthetic', 'esthetic', 'aesthetical'	1	concerning or characterized by an appreciation of beauty or good taste
69948	s	Synset('artistic.s.02')	['artistic']	1	satisfying aesthetic standards and sensibilities
70111	s	Synset('cosmetic.s.02')	['cosmetic', 'enhancive']	1	serving an aesthetic purpose in beautifying the body
70288	s	Synset('painterly.s.01')	['painterly']	1	having qualities unique to the art of painting
70427	s	Synset('sensuous.s.01')	['sensuous']	1	taking delight in beauty
70939	a	Synset('affected.a.01')	['affected']	1	acted upon; influenced
71142	s	Synset('impressed.s.01')	['impressed']	1	deeply or markedly affected or influenced
71242	s	Synset('smitten.s.01')	['smitten', 'stricken', 'struck']	1	(used in combination) affected by something overwhelming
71427	s	Synset('stage-struck.s.01')	['stage-struck']	1	infatuated with or enthralled by the theater especially the desire to act
71559	s	Synset('subject.s.03')	['subject']	1	likely to be affected by something
71739	s	Synset('taken.s.02')	['taken']	1	be affected with an indisposition
71897	s	Synset('wonder-struck.s.01	['wonder-struck']	1	affected by or overcome with wonder
73048	a	Synset('affected.a.02')	['affected', 'unnatural']	1	speaking or behaving in an artificial way to make an impression
73358	s	Synset('agonistic.s.03')	['agonistic', 'strained']	1	struggling for effect
73465	s	Synset('artificial.s.02')	['artificial', 'contrived', 'hokey', 'stil	1	artificially formal
73761	s	Synset('constrained.s.01')	['constrained', 'forced', 'strained']	1	lacking spontaneity; not natural
73935	s	Synset('elocutionary.s.02')	['elocutionary']	1	(used of style of speaking) overly embellished
74094	s	Synset('mannered.s.01')	['mannered']	1	having unnatural mannerisms
74216	s	Synset('plummy.s.02')	['plummy']	1	(of a voice) affectedly mellow and rich
75135	a	Synset('affirmative.a.01')	['affirmative', 'affirmatory']	1	affirming or giving assent
82766	s	Synset('battleful.s.01')	['battleful', 'bellicose', 'combative'	1	having or showing a ready disposition to fight
83003	s	Synset('competitive.s.03')	['competitive', 'militant']	1	showing a fighting disposition
83296	s	Synset('hard-hitting.s.02')	['hard-hitting', 'high-pressure']	1	aggressively and persistently persuasive

Fig. 13. Fragment of lex^{1}_{dis} positive list.

SYN ID	SYN POS	SYN	SYN TERMS	CONF	GLOSS
88055	s	Synset('jolted.s.01')	['jolted']	-1	bumped or shaken jerkily
88157	s	Synset('rippled.s.02')	['rippled', 'ruffled']	-1	shaken into waves or undulations as by wind
88328	s	Synset('seething.s.01')	['seething']	-1	in constant agitation
88545	s	Synset('stirred.s.03')	['stirred']	-1	set into a usually circular motion in order to mix or blend
89355	a	Synset('disagreeable.a.01')	['disagreeable']	-1	not to your liking
89550	s	Synset('annoying.s.01')	['annoying', 'bothersome', 'galling'	-1	causing irritation or annoyance
90219	s	Synset('harsh.s.06')	['harsh', 'abrasive']	-1	sharply disagreeable; rigorous
90408	s	Synset('nerve-racking.s.01'	['nerve-racking', 'nerve-wracking',	-1	extremely irritating to the nerves
90628	s	Synset('unsweet.s.02')	['unsweet']	-1	distasteful
101800	a	Synset('egoistic.a.01')	['egoistic', 'egoistical', 'egocentric',	-1	limited to or caring only about yourself and your own needs
104318	s	Synset('pushful.s.01')	['pushful', 'pushy']	-1	marked by aggressive ambition and energy and initiative
112628	a	Synset('synthetic.a.02')	['synthetic', 'synthetical']	-1	involving or of the nature of synthesis (combining separate elements to form a coherent wh
113818	a	Synset('angry.a.01')	['angry']	-1	feeling or showing anger
115494	s	Synset('indignant.s.01')	['indignant', 'incensed', 'outraged',	-1	angered at something unjust or wrong
117754	a	Synset('insentient.a.01')	['insentient', 'insensate']	-1	devoid of feeling and consciousness and animation
117961	s	Synset('unfeeling.s.02')	['unfeeling']	-1	devoid of feeling or sensation
133851	a	Synset('unappetizing.a.01')	['unappetizing', 'unappetising']	-1	not appetizing in appearance, aroma, or taste
140989	s	Synset('columnar.s.02')	['columnar']	-1	characterized by columns
147734	a	Synset('artful.a.02')	['artful']	-1	marked by skill in achieving a desired end especially with cunning or craft
148078	s	Synset('crafty.s.01')	['crafty', 'cunning', 'dodgy', 'foxy', '	-0.5	marked by skill in deception
149861	a	Synset('artless.a.02')	['artless']	-1	simple and natural; without cunning or deceit
151855	s	Synset('dumb.s.04')	['dumb', 'mute', 'silent']	-1	unable to speak because of hereditary deafness
152285	s	Synset('mute.s.01')	['mute', 'tongueless', 'unspoken', '	-1	expressed without speech
153898	a	Synset('ashamed.a.01')	['ashamed']	-0.333	feeling shame or guilt or embarrassment or remorse
154163	s	Synset('discredited.s.02')	['discredited', 'disgraced', 'dishono	-1	suffering shame
154270	s	Synset('embarrassed.s.02')	['embarrassed', 'humiliated', 'mort	-1	made to feel uncomfortable because of shame or wounded pride
154837	s	Synset('shamefaced.s.02')	['shamefaced', 'sheepish']	-1	showing a sense of shame
156440	s	Synset('cocky.s.01')	['cocky']	-1	overly self-confident or self-assertive
157268	s	Synset('reticent.s.03')	['reticent', 'self-effacing', 'retiring']	-1	reluctant to draw attention to yourself
157925	a	Synset('attached.a.03')	['attached', 'committed']	-0.333	associated in an exclusive sexual relationship
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Fig. 14. Fragment of lex^{1}_{dis} negative list.

5.1.3. Prominent empirical observations on subjectivity in the English language

Figs. 13 and 14 depict a fragment of the positive list and the negative list in the lexicon respectively. The six columns list the synset ID, part of speech, synset, synset member terms, confidence score, and gloss. This lexicon contains a total of 8101 positive synsets (2947 adjectives, 144 adverbs, 4050 nouns, and 960 verbs), and 3186 negative synsets (1457 adjectives, 88 adverbs, 1253 nouns, and 388 verbs). The conf values represent the measure of confidence that each synset belongs to its corresponding category. The lexicon can be fine-tuned based on the conf values, e.g. retaining only senses with conf values over a certain threshold to obtain higher precision, at the cost of a smaller-sized lexicon (e.g. retaining only synsets over a conf of > 0.75). The lex^{1}_{dis} sentiment lexicon is made publically available for research purposes[6].

Using the lex^{1}_{dis} generation model (the best-perfroming model in terms of accuracy), some interesting empirical observations are highlighted with regards to subjectivity in the English language. Since objectivity is the absence of subjectivity, and the assumption is made the WordNet represents the English language (117,659 synsets in total), it can be inferred that the remaining synsets in WordNet that have not been labeled as subjective (positive or negative) by the model, are automatically assumed to be objective. Recalling that lex^{1}_{dis} contains 11,287 subjective (8101 positive and 3186 negative) synsets, this amounts to 117,659 WordNet synsets – 11,287 subjective synsets = 106,372 objective synsets. With this, the following empirical observations are highlighted: (1) the English language is nearly 10% subjective, and 90% objective, and (2) among the subjective portion, the language leans 70% towards positivity, and 30% towards negativity.

Observation (1) is in line with Esuli and Sebastiani (2006) observation that English contains 10.45% words that have a subjective (positive or negative) score of at least 0.5, and that there are only a few words in English that are “unquestionably positive or negative”. This has also been demonstrated by Adreevskaia and Bergler (2006), who mention that only a small set of ‘core’ words are central to their respective sentiment categories, and have a strongly defined polarity.

Observation (2) demonstrates that English contains a significantly higher amount of positivity than negativity, and can be assumed to be an ‘optimistic language’. This is in agreement with Kloumann, Danforth, Harris, Bliss, and Dodds (2012), who empirically demonstrate a clear positive bias in an empirical investigation on the 10,000 most frequently used words in the English language.

5.2. Extrinsic evaluation results

Two extrinsic evaluation procedures are performed. The first one compares the best-performing generated sense-level lexicon (lex^{1}_{dis}) to other state-of-the-art sense-level lexicons (Section 5.2.1). The second one compares this lexicon (lex^{1}_{dis}) to a term-level version of the same lexicon (Section 5.2.2).

As mentioned earlier, four independent datasets are used for each evaluation procedure, which are the SLMRs, BRMSD, ERMSD and SERMSD datasets.

[6] http://www.ftsm.ukm.my/SOMRsentimentlexicon.zip

Table 7
Evaluation results for generated lexicon vs. prominent state-of-the-art lexicons on the SLMRs.

Lexicon used	P	R	F
Generated lex^1_{dis} lexicon	60.60	62.10	**61.34**
SentiWordNet 3.0	58.40	62.20	60.20
QWordNet	57.20	58.60	57.80
QWN-PPV	51.40	57.80	54.40

5.2.1. Comparison of generated lexicon vs. state-of-the-art lexicons

This section presents a comparison of the best performing (lex^1_{dis}) lexicon to prominent state-of-the-art sense-level lexicons, in a polarity classification task on the mentioned datasets. The P, R and F values are computed for each lexicon, for each of the datasets.

Table 7 presents the P, R and F values for the generated lexicon, QWN-PPV (San Vicente et al., 2014), SentiWordNet 3.0 (Baccianella et al. 2010), and QWordNet (Agerri & García-Serrano 2010), allowing for a side by side comparison of the performance of each lexicon in a sense-level sentiment classification task on the Stanford Large Movie Reviews Dataset (SLMRs).

According to Table 7, the generated lexicon outperforms all state-of-the-art sense-level sentiment lexicons on the SLMRs. The only exception is that SentiWordNet 3 has comparable R value to the generated lexicon (62.20 vs. 62.10 respectively). The underlying reason for this exception is because the SentiWordNet 3 lexicon generation model has a similar performance in its ability to retrieve polar senses. However, it is crucial to note that although it does a slightly better job at retrieving polar senses (i.e. slightly higher recall), its lower precision (58.40) demonstrates its inferior performance in the task of correctly matching those retrieved polar senses into their respectful categories of positive or negative.

The generated lexicon performs with an F-measure of 61.34, outperforming SentiWordNet 3 by 1.14%, QWordNet by 3.54% and QWN-PPV by 6.94%, in a binary sentiment classification task on the SLMRs, in a scenario in which all external parameters are fixed.

Table 8 presents the P, R and F values for the generated lexicon, QWN-PPV, SentiWordNet 3.0, and QWordNet, allowing for a side by side comparison of the performance of each lexicon in a sense-level sentiment classification task on the Book Reviews Multi-Domain Sentiment Dataset (BRMSD).

According to Table 8, the generated lexicon outperforms all state-of-the-art sense-level sentiment lexicons on the BRMSD. The generated lexicon performs with an F-measure of 59.36, outperforming SentiWordNet 3 by 5.22%, QWordNet by 6.15% and QWN-PPV by 8.69%, in a binary sentiment classification task on the BRMSD, in a scenario in which all external parameters are fixed.

Table 9 presents the P, R and F values for the generated lexicon, QWN-PPV, SentiWordNet 3.0, and QWordNet, allowing for a side by side comparison of the performance of each lexicon in a sense-level sentiment classification task on the Electronic Reviews Multi-Domain Sentiment Dataset (ERMSD).

According to Table 9, the generated lexicon outperforms all state-of-the-art sense-level sentiment lexicons on the ERMSD. The generated lexicon performs with an F-measure of 62.07, outperforming SentiWordNet 3 by 10.2%, QWordNet by 12.01% and QWN-PPV by 12.43%, in a binary sentiment classification task on the ERMSD, in a scenario in which all external parameters are fixed.

Table 10 presents the P, R and F values for the generated lexicon, QWN-PPV, SentiWordNet 3.0, and QWordNet, allowing for a side by side comparison of the performance of each lexicon in a sense-level sentiment classification task on the Sports Equipment Reviews Multi-Domain Sentiment Dataset (SERMSD).

According to Table 10, the generated lexicon outperforms all state-of-the-art sense-level sentiment lexicons on the SERMSD. The generated lexicon performs with an F-measure of 64.11, outperforming SentiWordNet 3 by 9.64%, QWordNet by 13.48% and QWN-PPV by 11.15%, in a binary sentiment classification task on the SERMSD, in a scenario in which all external parameters are fixed.

The results obtained in this evaluation procedure across four publically-available datasets of varying domains empirically demonstrate that the generated sense-level lexicon yields superior performance over employing related sense-level lexicons in a real-world sentiment classification task on full-text. This is illustrated in Fig. 15, whereby the F-measure achieved by the sense-level lexicon remains higher than the F-measure achieved by the related sense-level lexicons, across all datasets considered in this evaluation procedure.

Comparing the generated lexicon to the SentiWordNet 3 lexicon, the former outperforms the latter across all datasets considered. Although the difference in *F*-measure across all datasets is only by a small margin, the SentiWordNet 3 generated model requires a committee of eight supervised machine learning classifiers (SVMs and Roccio classification models), whereby each individual classifier performs a two-step classification task, as well as two time-consuming random walk steps (one for the positive class and one for the negative class). In sharp contrast, the proposed model is fully unsupervised and does not require any time-consuming supervised

Table 8
Evaluation results for generated lexicon vs. prominent state-of-the-art lexicons on the BRMSD.

Lexicon used	**P**	**R**	**F**
Generated lex^1_{dis} lexicon	59.10	59.63	**59.36**
SentiWordNet 3	53.04	55.29	54.14
QWordNet	52.80	53.63	53.21
QWN-PPV	50.21	51.14	50.67

Table 9
Evaluation results for generated lexicon vs. prominent state-of-the-art lexicons on the ERMSD.

Lexicon used	P	R	F
Generated lex^1_{dis} lexicon	62.12	62.03	**62.07**
SentiWordNet 3	51.71	52.04	51.87
QWordNet	49.96	50.16	50.06
QWN-PPV	49.59	49.68	49.64

Table 10
Evaluation results for generated lexicon vs. prominent state-of-the-art lexicons on the SERMSD.

Lexicon used	P	R	F
Generated lex^1_{dis} lexicon	64.09	64.15	**64.12**
SentiWordNet 3	53.92	55.05	54.48
QWordNet	50.59	50.69	50.64
QWN-PPV	51.53	54.5	52.97

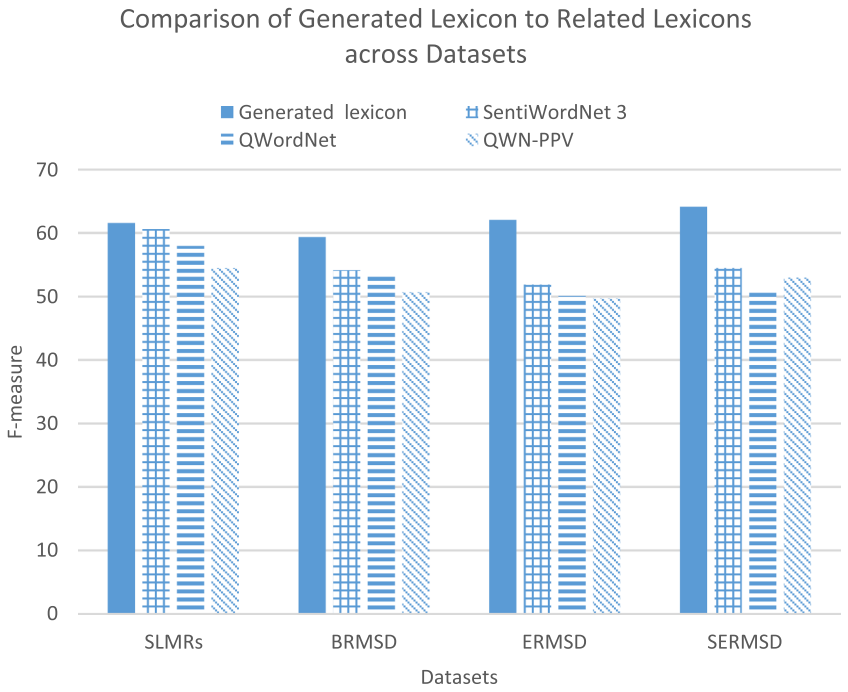

Fig. 15. Comparison of generated lexicon to related lexicons across datasets.

classifiers. Their model also employs two time-consuming random walk steps based on an in-gloss matching method which does not consider the contextual structure of the gloss information (e.g. negation words), whereas the proposed model is context-aware. Additionally, their model initiates with a total of 105 manually-labelled synsets, while the proposed model initiates with null seed sets. The only input into the model is one SOMR (e.g. dis~), and the model runs in a fully automated manner, from the initiation to its termination, without the reliance on any other human input or involvement.

Furthermore, applying the resultant generated lexicons side by side in this polarity classification task demonstrates that the proposed context-aware gloss expansion algorithm seems to have potentially addressed the limitation inherent in the semantic relations expansion algorithm by picking up remote polar senses in the network. They apply the semantic relations algorithm in their model, which, as pointed out earlier, fails to detect remote polar synsets beyond its reach, since the semantic relations become weaker as the path between the seed synset and the target synset increases. It is crucial to highlight that the proposed context-aware gloss expansion algorithm has successfully addressed this limitation, whereby in-gloss matching is able to pick up remote potentially subjective synsets that are overlooked by the semantic relations algorithm in the SentiWordNet 3 generation model. This is because the proposed context-aware gloss expansion algorithm relies on semantic information within glosses, and is not constrained to semantic distance within the semantic network.

Comparing the generated lexicon to the QWordNet lexicon, the former outperforms the latter across all datasets considered. This

is due to the fact that the QWordNet lexicon generation model has a conservative nature in its polarity classification technique, filtering out subtly-subjective terms as objective. In contrast, the proposed model treats the objective class with an equal priority as the positive and negative classes, only adding synsets to that class that fail to have any polar characteristics (e.g. are linked to many polar seed synsets in the semantic network, or are contained within the glosses of many polar seed synsets). Moreover, it fails to utilise gloss information during the categorisation process, while the proposed model employs the context-aware gloss expansion algorithm to effectively utilize gloss information to aid in the classification process. Finally, the model uses all semantic relations to traverse the graph, but some relations do not effectively preserve sentiment properties, e.g. the *attribute* relation of the synset happiness.n.01 is linked to both happy.a.01 and unhappy.a.01 at the same time. In contrast, the proposed model utilizes only relations in the semantic network that have been tested to reliably preserve sentiment polarity during propagation.

Comparing the generated lexicon to the QWN-PPV lexicon, the former outperforms the latter across all datasets considered. This is because, similar to the previously discussed QWordNet lexicon generation model, the QWN-PPV lexicon generation model also possesses similar limitations. First, it has a conservative nature in its polarity classification technique, filtering out subtly-subjective terms as objective. Second, the model uses all semantic relations to traverse the graph, even the relations that do not effectively preserve sentiment properties (e.g. attribute).

After comparing the generated lexicon to related lexicons across multiple datasets, it can be concluded that the generated lexicon achieves superior performance to existing state-of-the-art lexicons, in terms of the standard evaluation metrics, despite the fact that this lexicon is generated using a fully-unsupervised approach with no human involvement. Additionally, the generated lexicon demonstrates stable performance and robustness when applied on datasets with varying domains.

5.2.2. Term-level vs sense-level classification

The lex^{1}_{dis} lexicon is also used in a second evaluation procedure involving a comparison of the generated sense-level lexicon vs a term-level version of the same lexicon. The Precision (P), Recall (R) and F-measure (F) values achieved by the generated sense-level lexicon are compared to the values achieved by a term-level version of the same lexicon, on the four datasets considered. For a sense-disambiguated version of each dataset, the 'pywsd' word sense disambiguation module[7] for Python is used to map document terms to their corresponding synsets in WordNet.

Table 11 shows the P, R and F values achieved by the generated sense-level lexicon as compared to the values achieved by a term-level version of the same lexicon, on the SLMRs.

According to the table, on the SLMRs, the sense-level lexicon achieved a 3.34% higher F-measure compared to the term-level lexicon. These results demonstrate the contribution to overall performance this finer-grained lexicon has over its term-level counterpart, on a dataset in the movie reviews domain.

Table 12 shows the P, R and F values achieved by the generated sense-level lexicon as compared to the values achieved by a term-level version of the same lexicon, on the BRMSD.

According to the table, on the BRMSD, the sense-level lexicon achieved a 5.44% higher F-measure compared to the term-level lexicon. These results demonstrate the contribution to overall performance this finer-grained lexicon has over its term-level counterpart, on a dataset in the book reviews domain.

Table 13 shows the P, R and F values achieved by the generated sense-level lexicon as compared to the values achieved by a term-level version of the same lexicon, on the ERMSD.

According to the table, on the ERMSD dataset, the sense-level lexicon achieved a 4.02% higher F-measure compared to the term-level lexicon. These results demonstrate the contribution to overall performance this finer-grained lexicon has over its term-level counterpart, on a dataset in the electronic product reviews domain.

Table 14 shows the P, R and F values achieved by the generated sense-level lexicon as compared to the values achieved by a term-level version of the same lexicon, on the SERMSD.

According to the table, on the SERMSD, the sense-level lexicon achieved a 3.46% higher F-measure compared to the term-level lexicon. These results demonstrate the contribution to overall performance this finer-grained lexicon has over its term-level counterpart, on a dataset in the sports equipment reviews domain.

The results obtained in this evaluation procedure across four publically-available datasets of varying domains empirically demonstrate that employing a sense-level sentiment lexicon, and a pre-processing step of word sense disambiguation on the text to be classified, yields superior performance over employing a term-level sentiment lexicon for the sentiment classification task. This is illustrated in Fig. 16, whereby the F-measure achieved by the sense-level lexicon remains higher than the F-measure achieved by the term-level lexicon, across all datasets considered in this evaluation procedure.

This claim that a finer-grained sentiment lexicon composed of term senses would be of more benefit compared to a lexicon of terms alone, is in agreement with prior work (Akkaya et al. 2009; Wiebe & Mihalcea 2006), since the different senses of a particular term may possess varying sentiment polarity. For example, a sense-level lexicon is able to distinguish between the sense mean.a.02, as in "hateful", conveying negativity; and mean.n.01, as in "statistical average", conveying objectivity. A term-level lexicon, on the other hand, would fail to differentiate between the two different senses of the word, and incorrectly assign one 'static' polarity to the word, regardless of the intended sense used. It can be concluded that, practically, a pre-processing pass of word sense disambiguation can be run on the text to be classified, and the particular term senses in the text can be mapped to their corresponding senses in the

[7] The pywsd module is a standardised Python implementation for practical word sense disambiguation of words. The official pywsd module can be accessed at https://pypi.org/project/pywsd.

Table 11
Evaluation results for term-level and sense-level classification on SLMRs.

Lexicon used	P	R	F
Sense-level	60.60	62.10	61.34
Term-Level	54.65	61.80	58.00

Table 12
Evaluation results for term-level and sense-level classification on BRMSD.

Lexicon used	P	R	F
Sense-level	59.10	59.63	59.36
Term-Level	52.16	55.81	53.92

Table 13
Evaluation results for term-level and sense-level classification on ERMSD.

Lexicon used	P	R	F
Sense-level	62.12	62.03	62.07
Term-Level	56.33	59.88	58.05

Table 14
Evaluation results for term-level and sense-level classification on SERMSD.

Lexicon used	P	R	F
Sense-level	64.09	64.15	64.12
Term-Level	58.14	63.41	60.66

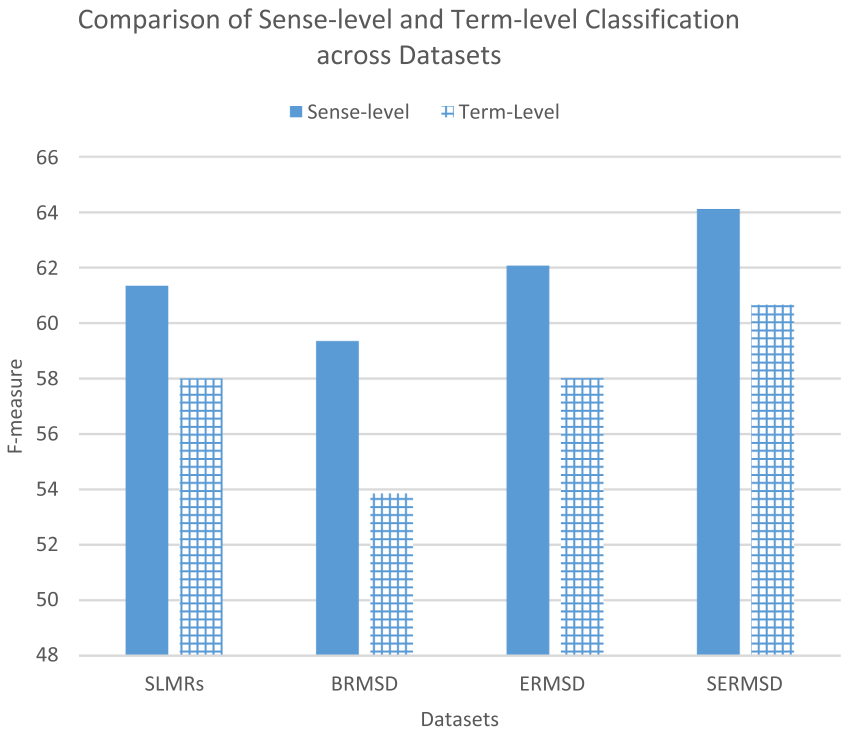

Fig. 16. Comparison of sense-level and term-level classification across datasets.

sentiment lexicon, for a more fine-grained sentiment classification compared to a naïve term-level classification, as demonstrated in this evaluation procedure.

6. Conclusion and future work

This work has proposed a sense-level sentiment lexicon generation model that derives the polarity of senses using null seed sets, dual-step context-aware in-gloss matching, and a fully-unsupervised sentiment categorization algorithm on the basis of the Network Theory. The results from intrinsic evaluation of the model against standardized gold standard benchmarks have demonstrated superior accuracy to state-of-the-art sense-level lexicon models. The proposed context-aware gloss expansion algorithm seems to have potentially addressed the limitation inherent in the semantic relations expansion algorithm, validated by its superior performance in comparison to state-of-the-art. The results from extrinsic evaluation in a real-world polarity classification task on four publically-available datasets of varying domains demonstrates its superior performance over existing sense-level lexicons across all datasets, its robustness in variable-domain scenarios, as well as its practical application in sense-level sentiment classification tasks on full-text.

The proposed model generates a general-purpose, domain-independent lexicon. Each term sense in the lexicon is assigned a prior polarity, which refers to its general, stereotypical, out-of-context polarity. Although it has been demonstrated that the resultant lexicon is robust and performs well across variable-domain datasets, the integration of a corpus during in domain-adaptation step can further fine-tune the sensitivity of the resultant lexicon to the target domain, topic or context considered, for improved results. For example, it should be able to detect that ‘predictable’ should be positive when discussing the stock market, but negative when mentioned in a movie or book review. The generated general-purpose lexicon may be used as a foundation in future work towards this research direction.

Additionally, we aim to employ a fully semantic approach in the quantification of the natural sentiment strength of the polar term senses generated by the model. In state-of-the-art, sentiment strength values rely purely on *statistical means*, and there is no *semantic mechanism* involved, leading to biased results. It can be argued that, semantically, this is an invalid representation of a sense's truly ‘natural’ sentiment strength, hence, the motivation to exploit the semantic, human-defined gloss information to quantify the natural sentiment strength of subjective term senses.

Acknowledgements

We would like to thank the Ministry of Higher Education for providing monetary assistance under the grant FRGS/2/2013/ICT02/UKM/02/1 and FRGS/1/2014/ICT02/UKM/01/1.

References

Abdaoui, A., Azé, J., Bringay, S., & Poncelet, P. (2017). Feel: a french expanded emotion lexicon. *Language Resources and Evaluation, 51*(3), 833–855.
Agerri, R., & García-Serrano, A. (2010). Q-WordNet: Extracting Polarity from WordNet Senses. *LREC*.
Akkaya, C., Wiebe, J., & Mihalcea, R. (2009). Subjectivity word sense disambiguation. *Proceedings of the conference on empirical methods in natural language processing* (pp. 190–199). .
Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information Processing & Management, 56*(2), 320–342.
Alqasemi, F., Abdelwahab, A., & Abdelkader, H. (2019). Constructing automatic domain-specific sentiment lexicon using KNN search via terms discrimination vectors. *International Journal of Computers and Applications, 41*(2), 129–139.
Al-Saffar, A., Awang, S., Tao, H., Omar, N., Al-Saiagh, W., & Al-Bared, M. (2018). Malay 1290 sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. *PLoS One, 14*(4), e0194852. https://doi.org/10.13140/RG.2.2.33420.72320.
Alshahrani, H. A., & Fong, A. C. (2018). Arabic domain-oriented sentiment lexicon construction using latent Dirichlet allocation. *Proceedings of the IEEE international conference on electro/information technology (EIT)* (pp. 0174–0180). IEEE.
Al-Thubaity, A., Alqahtani, Q., & Aljandal, A. (2018). Sentiment lexicon for sentiment analysis of Saudi dialect tweets. *Procedia computer science, 142*, 301–307.
Andreevskaia, A., & Bergler, S. (2006). Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. *Proceedings of the 11th conference of the European chapter of the association for computational linguistics*.
Asgari, E., Braune, F., Ringlstetter, C., & Mofrad, M.R. (2019). UniSent: Universal Adaptable Sentiment Lexica for 1000+ Languages. arXiv preprint arXiv:1904.09678.
Asghar, M. Z., Sattar, A., Khan, A., Ali, A., Masud Kundi, F., & Ahmad, S. (2019). Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. *Expert Systems,* e12397.
Assiri, A., Emam, A., & Al-Dossari, H. (2018). Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. *Journal of Information Science, 44*(2), 184–202.
Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *LREC, 10*(2010), 2200–2204.
Battistella, E. L. (1990). *Markedness: The evaluative superstructure of language.* SUNY Press.
Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G., & Reynar, J. (2008). Building a sentiment summarizer for local service reviews.
Bollegala, D., Weir, D., & Carroll, J. (2011). Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 1. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 132–141). Association for Computational Linguistics.
Burt, R. S. (1980). Models of network structure. *Annual review of sociology, 6*(1), 79–141.
Carrillo-de-Albornoz, J., Plaza, L., & Gervás, P. (2012). SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. *Proceedings of the LREC. 12. Proceedings of the LREC* (pp. 3562–3567).
Chaturvedi, I., Cambria, E., Welsch, R. E., & Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion, 44*, 65–77.
Councill, I. G., McDonald, R., & Velikovich, L. (2010). What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. *Proceedings of the workshop on negation and speculation in natural language processing* (pp. 51–59). Association for Computational Linguistics.
Darwich, M., Noah, S. A. M., & Omar, N. (2015). *Inducing a domain-independent sentiment lexicon in Malay. JAIST Symposium on Advance Science and Technology*. JAIST.
Darwich, M., Noah, S. A. M., & Omar, N. (2016). Automatically generating a sentiment lexicon for the Malay language. *Asia-Pacific Journal of Information Technology*

and Multimedia, 5(1), 49–69.
Darwich, M., Noah, S. A. M., & Omar, N. (2017). Minimally-supervised sentiment lexicon induction model: A case study of malay sentiment analysis. *Proceedings of the International Workshop on Multi-disciplinary Trends in Artificial Intelligence* (pp. 225–237). Springer.
Dehkharghani, R. (2018). A hybrid approach to generating adjective polarity lexicon and its application to turkish sentiment analysis. *International Journal of Modern Education and Computer Science, 10*(11), 11.
Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems, 94*, 65–76.
Dragut, E., Wang, H., Yu, C., Sistla, P., & Meng, W. (2012). Polarity consistency checking for sentiment dictionaries. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. 1. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers* (pp. 997–1005). Association for Computational Linguistics.
Ekinci, E., & Omurca, S.İ (2019). A new approach for a domain-independent Turkish sentiment seed lexicon compilation. *International Arab Journal of Information Technology, 5*, 1–11.
El-Beltagy, S. R. (2016). NileULex: a phrase and word level sentiment lexicon for Egyptian and modern standard Arabic. *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 2900–2905). .
Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *In LREC (Vol, 6*, 417–422.
Feng, S., Zhang, L., Li, B., Wang, D., Yu, G., & Wong, K. F. (2013). Is Twitter a better corpus for measuring sentiment similarity? *Proceedings of the conference on empirical methods in natural language processing* (pp. 897–902). .
Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications, 58*, 57–75.
Gatti, L., & Guerini, M. (2012). Assessing sentiment strength in words prior polarities. arXiv preprint arXiv:1212.4315.
Guellil, I., Adeel, A., Azouaou, F., & Hussain, A. (2018). Sentialg: Automated corpus annotation for algerian sentiment analysis. *Proceedings of the international conference on brain inspired cognitive systems* (pp. 557–567). Springer.
Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing. 2016. Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing* (pp. 595–). NIH Public Access.
Hassan, A., & Radev, D. (2010). Identifying text polarity using random walks. *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 395–403). Association for Computational Linguistics.
Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics* (pp. 174–181). Association for Computational Linguistics.
Huang, S., Niu, Z., & Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems, 56*, 191–200.
Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences, 30*(4), 330–338.
Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the eighth international AAAI conference on weblogs and social media*.
Ide, N. (2006). Making senses: Bootstrapping sense-tagged lists of semantically-related words. *Proceedings of the international conference on intelligent text processing and computational linguistics* (pp. 13–27). Springer.
Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing, 3*.
Kaity, M., & Balakrishnan, V. (2018). Building multilingual sentiment lexicons based on unlabelled corpus. *Proceedings of the Data Science Research Symposium* (pp. 10). .
Kaity, M., & Balakrishnan, V. (2019). An automatic non-English sentiment lexicon builder using unannotated corpus. *The Journal of Supercomputing, 75*(4), 2243–2268.
Kanayama, H., & Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. *Proceedings of the conference on empirical methods in natural language processing* (pp. 355–363). Association for Computational Linguistics.
Kandé, D., Camara, F., Ndiaye, S., & Guirassy, F. M. (2019). FWLSA-score: French and wolof lexicon-based for sentiment analysis. *Proceedings of the 5th international conference on information management (ICIM)* (pp. 215–220). IEEE.
Kannan, A. (2019). *Sentiment lexicon creation in tamil using hybrid techniques.* International Institute of Information Technology Hyderabad Doctoral dissertation.
Kimura, M., & Katsurai, M. (2017). Automatic construction of an emoji sentiment lexicon. *Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining 2017* (pp. 1033–1036). ACM.
Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., & Dodds, P. S. (2012). Positivity of the English language. *PloS One, 7*(1), e29484 Journal Article.
Koltsova, O. Y., Alexeeva, S., & Kolcov, S. (2016). An opinion word lexicon and a training dataset for russian sentiment analysis of social media. *Computational Linguistics and Intellectual Technologies: Materials of Dialogue, 2016*(Moscow), 277–287.
Kong, L., Li, C., Ge, J., Yang, Y., Zhang, F., & Luo, B. (2018). Construction of microblog-specific chinese sentiment lexicon based on representation learning. *Proceedings of the pacific rim international conference on artificial intelligence* (pp. 204–216). Springer.
Lehrer, A. (1974). Semantic fields and lexical structure.
Li, J., & Hovy, E. (2017). *Reflections on sentiment/opinion analysis. A practical guide to sentiment analysis*. Springer41–59.
Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press.
Liu, B. (2017). *Many facets of sentiment analysis. A practical guide to sentiment analysis*. Springer11–39.
Loukachevitch, N., & Levchik, A. (2016). Creating a general russian sentiment lexicon. *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 1171–1176). .
Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 1. Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 142–150). Association for Computational Linguistics.
Machado, M. T., Pardo, T. A., & Ruiz, E. E. S. (2018). Creating a Portuguese context sensitive lexicon for sentiment analysis. *Proceedings of the international conference on computational processing of the Portuguese language* (pp. 335–344). Springer.
Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16–32.
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235–244.
Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. *Proceedings of the conference on empirical methods in natural language*.
Mohammad, S., Salameh, M., & Kiritchenko, S. (2016). Sentiment lexicons for Arabic social media. *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC'16*, 33–37.
Morante, R., & Sporleder, C. (2012). Modality and negation: An introduction to the special issue. *Computational Linguistics, 38*(2), 223–260.
Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning (No. 47)*. University of Illinois press.
Osman, N. A., Noah, S. A. M., & Darwich, M. (2019). Contextual sentiment based recommender system to provide recommendation in the electronic products domain. *International Journal of Machine Learning and Computing, 9*(4), 425–431. https://doi.org/10.18178/ijmlc.2019.9.4.821.
Peng, W., & Park, D. H. (2011). Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. *Proceedings of the fifth international AAAI conference on weblogs and social media*.
Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: research challenges, datasets, and recent advances. arXiv preprint arXiv:1905.02947.

Rao, D., & Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. *Proceedings of the 12th conference of the European chapter of the association for computational linguistics* (pp. 675–682). Association for Computational Linguistics.
Rehman, Z. U., & Bajwa, I. S. (2016). Lexicon-based sentiment analysis for Urdu language. *Proceedings of the sixth international conference on innovative computing technology (INTECH)* (pp. 497–501). IEEE.
Rouces, J., Tahmasebi, N., Borin, L., & Eide, S. R. (2018). SenSALDO: Creating a sentiment lexicon for Swedish. *Proceedings of the eleventh international conference on language resources and evaluation (LREC-2018*.
Saif, H., Fernandez, M., Kastler, L., & Alani, H. (2017). Sentiment lexicon adaptation with context and semantics for the social web. *Semantic Web, 8*(5), 643–665.
San Vicente, I., Agerri, R., & Rigau, G. (2014). Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 88–97). Association for Computational Linguistics.
Saputra, F. T., & Nurhadryani, Y. (2018). Analysis of Indonesian Sentiments Using Indonesian Sentiment Lexicon by Considering Denial. *Proceedings of the international conference on advanced computer science and information systems (ICACSIS)* (pp. 361–366). IEEE.
Schneider, A., Male, J., Bhogadhi, S., & Dragut, E. (2018). DebugSL: An Interactive Tool for Debugging Sentiment Lexicons. *Proceedings of the conference of the North American chapter of the association for computational linguistics: Demonstrations* (pp. 36–40). .
Severyn, A., & Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015* (pp. 464–469). .
Song, M., Park, H., & Shin, K. S. (2019). Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean. *Information Processing & Management, 56*(3), 637–653.
Stone, P.J., Dunphy, D.C., & Smith, M.S. (1966). The general inquirer: A computer approach to content analysis.
Suktarachan, M. (2018). The development of semi-automatic sentiment lexicon construction tool for thai sentiment analysis. *Proceedings of the advances in natural language processing, intelligent informatics and smart technology: selected revised papers from the eleventh international symposium on natural language processing (SNLP-2016) and the first workshop in intelligent informatics and smart technology684. Proceedings of the advances in natural language processing, intelligent informatics and smart technology: selected revised papers from the eleventh international symposium on natural language processing (SNLP-2016) and the first workshop in intelligent informatics and smart technology* (pp. 97–). Springer 10-12 February 2016.
Taboada, M., Anthony, C., & Voll, K. D. (2006). Methods for creating semantic orientation dictionaries. *Proceedings of the LREC* (pp. 427–432). .
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307.
Tai, Y. J., & Kao, H. Y. (2013). Automatic domain-specific sentiment lexicon generation with label propagation. *Proceedings of the international conference on information integration and web-based applications & services* (pp. 53). ACM.
Tang, D., Wei, F., Qin, B., Liu, T., & Zhou, M. (2014). Coooolll: A deep learning system for twitter sentiment classification. *Proceedings of the 8th international workshop on semantic evaluation (SemEval)* (pp. 208–212). .
Thelwall, M. (2017). The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. *Proceedings of the Cyberemotions* (pp. 119–134). Springer.
Tran, T. K., & Phan, T. T. (2018). A hybrid approach for building a Vietnamese sentiment dictionary. *Journal of Intelligent & Fuzzy Systems, (Preprint,* 1–12.
Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS), 21*(4), 315–346.
Velikovich, L., Blair-Goldensohn, S., Hannan, K., & McDonald, R. (2010). The viability of web-derived polarity lexicons. *Proceedings of the human language technologies: the 2010 annual conference of the north American chapter of the association for computational linguistics* (pp. 777–785). Association for Computational Linguistics.
Vo, D. T., & Zhang, Y. (2016). Don't count, predict! an automatic approach to learning sentiment lexicons for short text. *Proceedings of the 54th annual meeting of the association for computational linguistics. 2. Proceedings of the 54th annual meeting of the association for computational linguistics* (pp. 219–224). Short Papers.
Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. *Proceedings of the thirty-first AAAI conference on artificial intelligence*.
Weichselbraun, A., Gindl, S., & Scharl, A. (2011). Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 1053–1060). ACM.
Wiebe, J., & Mihalcea, R. (2006). Word sense and subjectivity. *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 1065–1072). .
Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. *Proceedings of the workshop on negation and speculation in natural language processing* (pp. 60–68). .
Williams, G. K., & Anand, S. S. (2009). Predicting the polarity strength of adjectives using wordnet. *Proceedings of the third international AAAI conference on weblogs and social media*.
Wu, F., Huang, Y., Song, Y., & Liu, S. (2016). Towards building a high-quality microblog-specific Chinese sentiment lexicon. *Decision Support Systems, 87*, 39–49.
Wu, L., Morstatter, F., & Liu, H. (2018). SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation, 52*(3), 839–852.
Wu, S., Wu, F., Chang, Y., Wu, C., & Huang, Y. (2019). Automatic construction of target-specific sentiment lexicon. *Expert Systems with Applications, 116*, 285–298.
Xing, F. Z., Pallucchini, F., & Cambria, E. (2019). Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management, 56*(3), 554–564.
Xu, T., Peng, Q., & Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Systems, 35*, 279–289.
Yang, C., Zhang, H., Jiang, B., & Li, K. (2019). Aspect-based sentiment analysis with alternating coattention networks. *Information Processing & Management, 56*(3), 463–478.
Zhang, S., Wei, Z., Wang, Y., & Liao, T. (2018). Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future Generation Computer Systems, 81*, 395–403.
Zhu, X., Guo, H., Mohammad, S., & Kiritchenko, S. (2014). An empirical study on the effect of negation words on sentiment. *Proceedings of the 52nd annual meeting of the association for computational linguistics. 1. Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 304–313). Long Papers.
Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 440-447).