SentiFul: Generating a Reliable Lexicon for Sentiment Analysis

Alena Neviarouskaya
University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan
lena@mi.ci.i.u-tokyo.ac.jp

Helmut Prendinger
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430, Japan
helmut@nii.ac.jp

Mitsuru Ishizuka
University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan
ishizuka@i.u-tokyo.ac.jp

Abstract

The main drawback of any lexicon-based sentiment analysis system is the lack of scalability. Thus, in this paper, we will describe methods to automatically generate and score a new sentiment lexicon, called SentiFul, and expand it through direct synonymy relations and morphologic modifications with known lexical units. We propose to distinguish four types of affixes (used to derive new words) depending on the role they play with regard to sentiment features: propagating, reversing, intensifying, and weakening.

1. Introduction and background

Sentiment analysis is nowadays a rapidly developing field with a variety of emerging approaches targeting the recognition of sentiment reflected in written language. Sentiment-related information can be encoded *lexically* within the actual words of the sentence, *syntactically* by means of subordinate clauses, and *morphologically* through changes in attitudinal shades of word meaning using suffixes (especially, in languages with rich inflectional system, such as Russian or Italian) [1].

Methods for extracting and annotating subjective terms include machine learning approaches examining the conjunction relations between adjectives [2], clustering adjectives according to distributional similarity based on a small amount of annotated seed words [3], pattern-bootstrapping algorithm to extract nouns [4], consideration of web-based mutual information to extract adjectives [5], and morphosyllabic sentiment tagging [6]. A useful sentiment lexicon would contain assignments of polarity orientation (positive and negative), and also the strength of sentiment or, in some cases, the degree of centrality to the sentiment category. To determine the word-level strength of sentiment, Latent Semantic Analysis [7], pointwise mutual information technique [7, 8], and methods employing WordNet structure relations [9, 10, 11] were proposed.

Most lexicon-based systems for sentiment analysis face the difficulty of assigning the sentiment scores to words that are not available in their databases. To deal with limitation in lexicon coverage, in this work, we will propose methods to automatically build and expand the subjectivity lexicon represented by sentiment-conveying words, which are annotated by sentiment polarity, polarity scores and weights. Although many researchers already attempted to extract and score new words through synonymy and antonymy relations, derivation of new sentiment lexemes by manipulation with morphological structure of words was not well explored. To our knowledge, the only work employing morphological analysis for sentiment tagging of unknown words is [6] (new word is transformed and compared with known sentiment lemmas and affixes). In our work, we approach the problem from the opposite direction: based on sentiment-scored lemmas and types of affixes, new words are automatically built and scored.

2. Building the lexicon of sentiment

2.1. Generating the core of sentiment lexicon

The first step in building the lexicon of sentiment-conveying terms involves the collection of relevant content part-of-speech words (adjectives, adverbs, nouns, and verbs), and the assignment of prior polarity scores (positivity score and negativity score) to each lexical unit. By "sentiment polarity score" we mean the strength or degree of intensity of sentiment. In our work, for both opposite valences, the bounds of the polarity score are 0.0 (indicating the absence of given orientation of sentiment) and 1.0 (the utmost value).

For the generation of the core of sentiment lexicon, we employ Affect database [12], which contains in total 2438 direct and indirect emotion-related entries: 918 adjectives (e.g., '*euphoric*', '*hostile*'), 243 adverbs (e.g., '*luckily*', '*miserably*'), 900 nouns (e.g., '*fright*', '*mercy*'), and 377 verbs (e.g., '*reward*', '*blame*'). The affective features of each distinct word in this database are encoded using nine emotions ('anger', 'disgust', 'fear', 'guilt', 'interest', 'joy', 'sadness', 'shame', and 'surprise'), and are represented as a vector of emotional state intensities that range from 0.0 to 1.0. Using emotional vectors, we interpreted the sentiment of Affect database entries by means of polarity scores and polarity weights. We considered three emotions ('interest', 'joy', and 'surprise') as having mainly positive orientation, and six emotions ('anger', 'disgust', 'fear', 'guilt', 'sadness', and 'shame') as negatively-valenced.

Positivity and negativity scores were calculated using

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

Table 1: Examples of words with sentiment annotations from SentiFul.

Affective word	POS	Non-zero-intensity emotions from Affect database emotional vector	Polarity scores		Polarity weights	
			Pos_score	***Neg_score***	***Pos_weight***	***Neg_weight***
tremendous	adjective	‘surprise:1.0’, ‘joy:0.5’, ‘fear:0.1’	0.75	0.1	0.67	0.33
pensively	adverb	‘sadness:0.2’, ‘interest:0.1’	0.1	0.2	0.5	0.5
success	noun	‘joy:0.9’, ‘interest:0.6’, ‘surprise:0.5’	0.67	0.0	1.0	0.0
regret	verb	‘guilt:0.2’, ‘sadness:0.1’	0.0	0.15	0.0	1.0

Eq. 1 and Eq. 2. Based on the Eq. 3 and Eq. 4, we derived the polarity weights.

$$Pos\_score = \left[\frac{\sum_{i=1}^{pos} Intensity(i)}{pos} \right], \quad (1)$$

$$Neg\_score = \left[\frac{\sum_{i=1}^{neg} Intensity(i)}{neg} \right], \quad (2)$$

$$Pos\_weight = \left[\frac{pos}{pos + neg} \right], \quad (3)$$

$$Neg\_weight = \left[\frac{neg}{pos + neg} \right], \quad (4)$$

where *Intensity* is intensity value of corresponding emotion in emotional vector; *pos* (*neg*) is the number of positive (negative) emotions having *Intensity*>0.0 in emotional vector, respectively.

We named our sentiment database as “SentiFul”. Some examples of SentiFul entries are listed in Table 1.

The main drawback of a sentiment analysis approach, which is purely relying on lexicon of sentiment-conveying terms, is the lack of scalability, since the recall of the lexical method depends on the coverage of the database used. Thus, to expand SentiFul, we first investigated the possibility to take advantage of sense-level scores from SentiWordNet [11].

2.2. Examining the SentiWordNet

SentiWordNet was developed based on WordNet [13] synsets comprised from synonymous terms. Motivated by the assumption that ‘*different senses of the same term may have different opinion-related properties*’, Esuli and Sebastiani [11] developed a method employing eight ternary classifiers and quantitatively analyzing the glosses associated with synsets. Three numerical scores (*Obj*(*s*), *Pos*(*s*), and *Neg*(*s*), which range from 0.0 to 1.0 and in sum equal to 1.0), characterizing to what degree the terms included in a synset are objective, positive, and negative, were automatically determined based on the proportion of classifiers assigning the corresponding label to the synset.

The question ‘*How reliable SentiWordNet is?*’ arouse at the very beginning of its exploration, just after analyzing the scores of synsets that include adjective ‘*happy*’ (Table 2). Three out of six synsets are characterized by negativity predominance (*Neg*(*s*) is greater than both *Pos*(*s*) and *Obj*(*s*)); in two synsets the scores of positivity prevail (*Pos*(*s*) is greater than both *Neg*(*s*) and *Obj*(*s*)); and one synset is completely objective (*Obj*(*s*)=1.0) in SentiWordNet. A sentiment analysis system employing sense disambiguation algorithm might obtain counter-intuitive results on the sentence ‘*Those were happiest days, I never felt such elation!*’, if scores for {happy(5), euphoric(1)} synset would be considered.

Let us now turn to the analysis of possibilities to extend SentiFul lexicon using SentiWordNet. As in SentiFul we restricted polarity scores and polarity weights to distinct lexemes (sentiment features of different senses of a term are unified), we considered two approaches to derive scores for each lexeme from SentiWordNet: (1) Method ‘FS’: take *Pos*(*s*), *Neg*(*s*), and *Obj*(*s*) scores of first synset for each lemma in SentiWordNet; (2) Method ‘*UNI*’: estimate unified positivity and negativity scores for each lemma in SentiWordNet using Eq. 5 and Eq. 6; and derive weights of positivity, negativity, and objectivity based on Eq. 7, Eq. 8, and Eq. 9. As there are synsets where *Pos*(*s*)=*Neg*(*s*)>0, all weights need to be normalized.

$$Uni\_Pos\_score = \left[\frac{\sum_{i=1}^{pos} Pos(s)(i)}{pos} \right], \quad (5)$$

Table 2: SentiWordNet scores for synsets containing adjective ‘happy’.

Synset with corresponding sense	***Pos*(*s*)**	***Neg*(*s*)**	***Obj*(*s*)**
{happy(2), pleased(3)}: experiencing pleasure or joy; ‘*happy you are here*’; ‘*pleased with the good news*’	0.0	**0.75**	0.25
{happy(3), felicitous(2)}: marked by good fortune; ‘*a felicitous life*’; ‘*a happy outcome*’	**0.875**	0.0	0.125
{happy(4)}: satisfied; enjoying well-being and contentment; ‘*felt content with her lot*’; ‘*quite happy to let things go on as they are*’	0.0	**0.75**	0.25
{happy(5), euphoric(1)}: exaggerated feeling of well-being or elation	0.125	**0.5**	0.375
{happy(6), well-chosen(1)}: well expressed and to the point; ‘*a happy turn of phrase*’; ‘*a few well-chosen words*’; ‘*a felicitous comment*’	0.0	0.0	**1.0**
{happy(1)}: enjoying or showing or marked by joy or pleasure or good fortune; ‘*a happy smile*’; ‘*spent many happy days on the beach*’; ‘*a happy marriage*’	**0.625**	0.25	0.125

$$Uni\_Neg\_score = \left[\frac{\sum_{i=1}^{neg} Neg(s)(i)}{neg} \right], \qquad (6)$$

$$Pos\_weight = \left[\frac{pos}{senses} \right], \qquad (7)$$

$$Neg\_weight = \left[\frac{neg}{senses} \right], \qquad (8)$$

$$Obj\_weight = \left[\frac{obj}{senses} \right], \qquad (9)$$

where *pos* is the number of lemma senses having *Pos*(*s*)(*i*)>=*Neg*(*s*)(*i*) and *Pos*(*s*)(*i*)>0; *neg* is number of lemma senses having *Neg*(*s*)(*i*)>=*Pos*(*s*)(*i*) and *Neg*(*s*)(*i*)>0; *obj* is number of lemma senses having *Obj*(*s*)(*i*)=1; *senses* is a total number of lemma synsets.

Using '*FS*' and '*UNI*' methods, we obtained scores for all 152050 distinct lemmas in SentiWordNet. In particular, total numbers of distinct lemmas having either *Obj*(*s*)<=0.5 (from '*FS*') or *Obj_weight*<=0.5 (from '*UNI*') are 14918 and 37414, respectively. In order to evaluate the appropriateness of scores derived from SentiWordNet, we created a 'gold standard' based on SentiFul (originating from manually annotated Affect database) entries and their scores. For the 'gold standard' we considered only those SentiFul entries that also occur in SentiWordNet: 750 adjectives, 237 adverbs, 894 nouns, 372 verbs. The evaluation was based on the comparison of valence of dominant score derived from SentiWordNet with the valence of dominant score from SentiFul 'gold standard'.

The rule for determination of valence of dominant score for a lemma in 'gold standard' is: if *Pos_score*>= *Neg_score* & *Pos_weight*>*Neg_weight* => positive else if *Pos_score*>*Neg_score* & *Pos_weight*=*Neg_weight* => positive else if *Neg_score*>=*Pos_score* & *Neg_weight*> *Pos_weight* => negative else if *Neg_score*>*Pos_score* & *Neg_weight*=*Pos_weight* => negative else if *Pos_score*= *Neg_score* & *Pos_weight*=*Neg_weight* => random else if *Pos_score*>*Neg_score* => positive else negative.

To obtain valence of dominant score within scores derived from SentiWordNet using '*FS*' and '*UNI*' methods, we propose four ways:

1. '*FS_strength*' (disregarding *Obj*(*s*)): if *Pos*(*s*)> *Neg*(*s*) => positive else if *Neg*(*s*)>*Pos*(*s*) => negative else if *Pos*(*s*)=*Neg*(*s*)=0.0 => neutral else random.
2. '*FS_obj*': if *Obj*(*s*)>0.5 => neutral else if *Pos*(*s*)> *Neg*(*s*) => positive else if *Neg*(*s*)>*Pos*(*s*) => negative else random.
3. '*UNI_strength*' (disregarding *Obj_weight*): if *Uni_Pos_score*>*Uni_Neg_score* => positive else if *Uni_Neg_score*>*Uni_Pos_score* => negative else if *Uni_Pos_score*=*Uni_Neg_score*=0.0 => neutral else random.
4. '*UNI_weight*': if *Obj_weight*>0.5 => neutral else if *Uni_Pos_score*>=*Uni_Neg_score* & *Pos_weight*> *Neg_weight* => positive else if *Uni_Pos_score*> *Uni_Neg_score* & *Pos_weight*=*Neg_weight* => positive else if *Uni_Neg_score*>=*Uni_Pos_score* & *Neg_ weight*>*Pos_weight* => negative else if *Uni_Neg_score*> *Uni_Pos_score* & *Neg_weight*=*Pos_ weight* => negative else if *Uni_Pos_score*=*Uni_Neg_ score* & *Pos_weight*= *Neg_weight* => random else if *Uni_Pos_score*> *Uni_Neg_score* => positive else negative.

Table 3 includes some examples of obtained results. The results of the evaluation of different methods for obtaining scores for adjectives, adverbs, nouns, and verbs based on SentiWordNet are displayed in Figure 1. As seen from the diagrams, more accurate scores were obtained for adjectives in comparison with other parts of speech, and the worst results were obtained for scoring

Table 3: Examples of the comparison of results from different methods with 'gold standard'.

Lemma (POS)	Method	*Pos_score*	*Neg_score*	*Pos_weight*	*Neg_weight*	Dominant	Result
weakness (noun)	SentiWordNet sense #1	0.0	0.125				
	SentiWordNet sense #2	0.125	0.625				
	SentiWordNet sense #3	0.0	0.375				
	SentiWordNet sense #4	0.5	0.125				
	SentiWordNet sense #5	0.0	0.875				
	SentiFul 'gold standard'	**0.0**	**0.2**	**0.0**	**1.0**	**negative**	
	'*FS_strength*'	0.0	0.125	-	-	negative	**hit**
	'*FS_obj*'	0.0	0.125	-	-	neutral	**neutral no hit**
	'*UNI_strength*'	0.5	0.5	0.2	0.8	random	**random hit**
	'*UNI_weight*'	0.5	0.5	0.2	0.8	negative	**hit**
congratulate (verb)	SentiWordNet sense #1	0.25	0.125				
	SentiWordNet sense #2	0.0	0.125				
	SentiWordNet sense #3	0.0	0.375				
	SentiWordNet sense #4	0.0	0.5				
	SentiFul 'gold standard'	**0.4**	**0.0**	**1.0**	**0.0**	**positive**	
	'*FS_strength*'	0.25	0.125	-	-	positive	**hit**
	'*FS_obj*'	0.25	0.125	-	-	neutral	**neutral no hit**
	'*UNI_strength*'	0.25	0.333	0.25	0.75	negative	**no hit**
	'*UNI_weight*'	0.25	0.333	0.25	0.75	negative	**no hit**

the verbs. The ‘*UNI*’ method performed better than the method based on consideration of scores of the first synset in SentiWordNet (‘*FS*’ method). The results we obtained while examining SentiWordNet were not satisfying, and we decided to seek other ways for extension of the SentiFul lexicon.

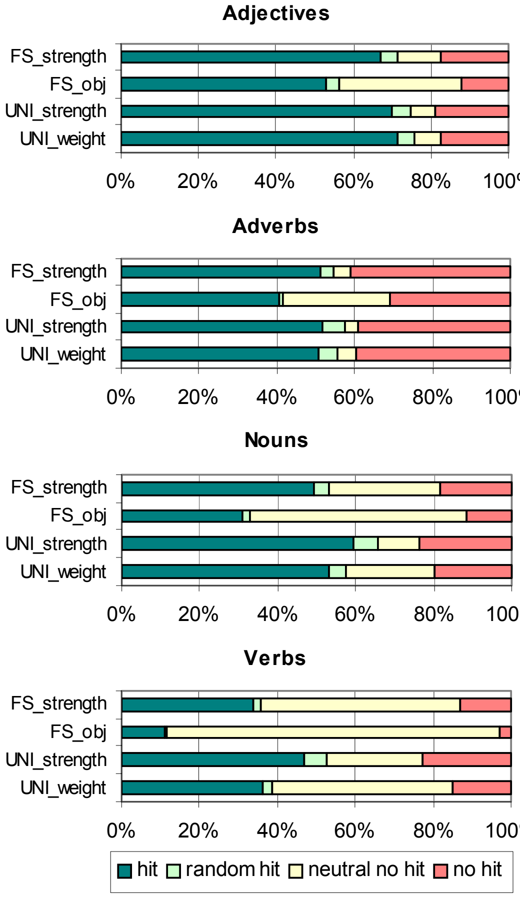

Figure 1: Accuracy of different methods for obtaining scores based on SentiWordNet.

2.3. Finding new lexical units through synonymy relation

To find new sentiment-related words, the most direct way is to derive them through the synonymy relation with known lexemes. Undoubtedly, the deep meaning of any lexical unit is unique. However, we can take advantage of considering the pairs of words, which have similar senses, while assigning the sentiment scores to them. The process of finding and scoring new words through a synonymy relation consists of three main steps, which are applied to adjectives, adverbs, nouns, and verbs independently.

Step 1. Given a word from SentiFul, we derive all related synsets found in WordNet. For example, four synsets were found for verb ‘*congratulate*’: {‘*compliment*’, ‘*congratulate*’}, {‘*congratulate*’, ‘*felicitate*’}, {‘*pride*’, ‘*plume*’, ‘*congratulate*’}, and {‘*preen*’, ‘*congratulate*’}.

Step 2. In each multiple-word synset from the previous step, we retrieve words that are already included in SentiFul, then estimate averages of scores and weights within synsets that have new terms, and finally assign these values to remaining words within corresponding synset. For the above example, all synonyms of the verb ‘*congratulate*’, except ‘*compliment*’ in first synset and ‘*felicitate*’ in second synset, are already in SentiFul. Therefore, scores of ‘*congratulate*’ (*Pos_score*=0.4, *Neg_score*=0.0, *Pos_weight*=1.0, and *Neg_weight*=0.0) are propagated to ‘*complement*’ and ‘*felicitate*’. In the case verb ‘*pride*’ from third synset was new for SentiFul, we would take the averages of polarity scores and averages of weights of both ‘*plume*’ and ‘*congratulate*’.

Step 3. After *Step 1* and *Step 2* are completed for all original SentiFul entries (we consider only their direct synonyms), we eliminate duplicates of new words, as they can obtain assignments from different synsets derived using different words from SentiFul, and estimate their new scores as averages of assignments of duplicates.

Relying on direct synonymy relations, we automatically extracted 4190 new words from WordNet (see examples in Table 4): 1122 adjectives, 107 adverbs, 1731 nouns, and 1230 verbs. We decided not to iterate the above procedure on these new words, because non-direct synonyms are not necessarily carrying similar sentiment features as original concepts (e.g., ‘*healthy*’-‘*intelligent*’-‘*thinking*’).

Table 4: Examples of newly derived words based on direct synonymy relations.

POS	Lemma	***Pos_score / Neg_score***	***Pos_weight / Neg_weight***
adjective	*appealing*	0.333 / 0.033	0.833 / 0.167
	barbarous	0.0 / 0.625	0.0 / 1.0
	confounded	0.1 / 0.2	0.167 / 0.833
adverb	*advantageously*	0.3 / 0.0	1.0 / 0.0
	frightfully	0.0 / 0.95	0.0 / 1.0
	poorly	0.0 / 0.334	0.0 / 1.0
noun	*authority*	0.383 / 0.05	0.875 / 0.125
	defect	0.0 / 0.6	0.0 / 1.0
	impetuosity	0.65 / 0.65	0.5 / 0.5
verb	*exhaust*	0.2 / 0.375	0.167 / 0.834
	glorify	0.3 / 0.0	1.0 / 0.0
	privilege	0.2 / 0.0	1.0 / 0.0

2.4. Method to derive and score morphologically modified words

We are proposing to expand our SentiFul lexicon through manipulations with morphological structure of known lemmas that result in a formation of new lexical units. Adjectives, adverbs, nouns, and verbs form open classes, whereby membership is indefinite and unlimited [14]. We can easily form new words playing with bases and affixes. Derivation is a process responsible for building new lexemes, either by adding derivational prefixes (attachments to the front of the base) or suffixes (attachments to the end of the base). Suffixes typically have less specific meanings than prefixes. The main contribution to meaning of many suffixes is that which follows from a change of grammatical class.

We distinguish four types of affixes depending on the role they play with regard to sentiment features:

1. *Propagating* affixes preserve sentiment features of the original lexeme and propagate them to newly derived lexical unit (e.g., ‘*en-*’+‘*rich*’=>‘*enrich*’, ‘*harmony*’+ ‘*-ous*’=>‘*harmonious*’, ‘*scary*’+‘*-fy*’=>‘*scarify*’).

2. *Reversing* affixes change the orientation of sentiment features of the original lexeme (e.g., ‘*dis-*’+ ‘*honest*’=>‘*dishonest*’, ‘*harm*’+‘*-less*’=>‘*harmless*’).

3. *Intensifying* affixes increase the strength of sentiment features of the original lexeme (e.g., ‘*super-*’+ ‘*hero*’=>‘*superhero*’, ‘*over-*’+‘*awe*’=>‘*overawe*’).

4. *Weakening* affixes decrease the strength of sentiment features of the original lexeme (e.g., ‘*semi-*’+‘*sweet*’=> ‘*semisweet*’).

Table 5 summarizes our classification with respect to type of an affix, class of a base lexeme (*a* stands for adjective, *adv* for adverb, *n* for noun, and *v* for verb), and class of a newly formed word.

Our algorithm for building new words receives the following parameters: class of the base word, class (prefix or suffix) and type of the affix, affix, and the class of derived word. The processing is as follows: (1) given the class of the base word, the system successively extracts each corresponding lemma from SentiFul and its sentiment-related scores, (2) depending on the affix class, affix is attached either to the front or to the end of the lemma to form new word, (3) given the class of derived word and newly formed word itself, SentiFul is scanned on the presence of this lemma, and if the result is positive, this lemma is not considered for inclusion, else, WordNet is examined on the availability of this lemma, and if this word exists, it is considered for future inclusion to SentiFul along with sentiment-related scores. Based on the type of the affix and sentiment-related scores of original word, scoring function assigns polarity scores and weights to the derived word. In the case of *Propagating* affix, original scores and weights are transferred to the new word invariably. The original *Pos_score* and *Neg_score* trade their places (same procedure for weights) in case of *Reversing* affix. If affix belongs to *Intensifying* or *Weakening* type, the original scores are multiplied by 2.0 or 0.5, respectively.

In order to properly treat attachment of suffixes to base lexemes, we apply next rules:

1. Replace lexeme ending ‘*f*’ (except the case of ‘*ff*’) by ‘*v*’ if suffix starts with ‘*a*/*e*/*i*/*o*/*u*/*y*’.

2. Replace lexeme ending ‘*fe*’ (except the case of ‘*ffe*’) by ‘*v*’ if suffix starts with ‘*a*/*e*/*i*/*o*/*u*/*y*’.

3. Remove lexeme ending ‘*y*’ if suffix starts with ‘*i*’.

4. Replace lexeme ending ‘*y*’, which follows the consonant, by ‘*i*’.

5. Remove (noun or adjective) lexeme ending ‘*t*’ or ‘*te*’ before suffix ‘*cy*’.

Table 5: Our classification of affixes attached to a base lexeme to form new word.

Type of affix	**Prefix (+*class of base lexeme*); (*class of base lexeme*+) suffix**
Adjective formation	
Propagating	pro- (+*a*); (*a*+) -ish; (*v*+) {-able, -ant, -ent, -ible, -ing}; (*n*+) {-al, -en, -ful, -ic, -like, -type, -y}; (*v*/*n*+) {-ate, -ed, -ive, -ous}
Reversing	{a-, ab-, an-, anti-, contra-, counter-, de-, dis-, dys-, il-, im-, in-, ir-, mal-, mis-, non-, pseudo-, un-, under-} (+*a*); (*n*+) -less
Intensifying	{extra-, hyper-, mega-, super-, ultra-} (+*a*)
Weakening	semi- (+*a*)
Adverb formation	
Propagating	pro- (+*adv*); (*a*+) -ly; (*n*+) {-wise, -wards}
Reversing	{a-, ab-, an-, anti-, contra-, counter-, de-, dis-, dys-, il-, im-, in-, ir-, mal-, mis-, non-, pseudo-, un-, under-} (+*adv*);
Intensifying	{extra-, hyper-, mega-, super-, ultra-} (+*adv*)
Weakening	semi- (+*adv*)
Noun formation	
Propagating	{neo-, re-} (+*n*); (*v*+) {-age, -al, -ant, -ation, -ent, -ication, -ification, -ion, -ment, -sion, -tion, -ure}; (*a*+) {-ity, -ness}; (*n*+) {-ful, ist, -ship}; (*v*/*a*+) {-ance, -ence, -ee}; (*v*/*n*+) {-er, -ing, -or}; (*a*/*n*+) {-cy, -dom, -hood}; (*v*/*n*/*a*) {-ery, -ry}
Reversing	{anti-, counter-, dis-, dys-, in-, mal-, mis-, non-, pseudo-, under-} (+*n*)
Intensifying	{arch-, hyper-, mega-, super-, ultra-} (+*n*)
Weakening	{mini-, semi-} (+*n*); (n+) {-ette, -let}
Verb formation	
Propagating	{be-, co-, fore-, inter-, pre-, pro-, re-, trans-} (+*v*); {em-, en-} (+*n*/*a*); (*n*/*a*+) {-ate, -en, -fy, -ify, -ise, -ize}
Reversing	{de-, dis-, dys-, mis-, un-, under-} (+*v*)
Intensifying	{out-, over-} (+*v*)

6. Remove lexeme ending ‘*e*’ if suffix starts with ‘*a*/*e*/*i*/*o*/*u*/*y*’.

7. Double lexeme ending ‘*b*/*d*/*f*/*g*/*l*/*m*/*n*/*p*/*r*/*s*/*t*/*v*/*z*’, which follows the vowel preceded by consonant, if suffix starts with ‘*a*/*e*/*i*/*o*/*u*/*y*’.

Using this morphologically inspired method, we automatically derived and scored 4029 new words (see examples in Table 6): 1405 adjectives, 484 adverbs, 1800 nouns, and 340 verbs.

Table 6: Examples of morphologically modified words.

POS	**Lemma**	***Pos_score* / *Neg_score***	***Pos_weight* / *Neg_weight***
adjective	*lovable*	0.85 / 0.0	1.0 / 0.0
	reproachful	0.0 / 0.625	0.0 / 1.0
adverb	*proficiently*	0.3 / 0.0	1.0 /0.0
noun	*spoilage*	0.133 / 0.3	0.167 / 0.833
verb	*beautify*	0.45 / 0.0	1.0 / 0.0

The *Propagating* type of affixes proved to be the most frequent and efficient in building words of all content parts of speech (Figure 2). The *Reversing* type of affixes played also significant role in the derivation process for adjectives and adverbs, while *Intensifying* affixes brought noticeable effect only in building new verbs.

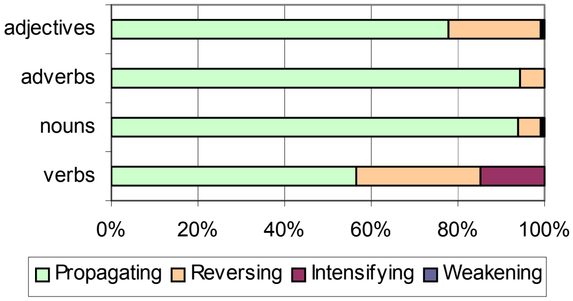

Figure 2: Percentage distribution of words derived by means of different affix types.

The block diagram shown in Figure 3 indicates that adjectives, adverbs, and nouns were mainly derived by means of suffixes; on the other hand, prefixes dominated in the case of verbs. The most productive affixes to form new words are listed in Table 7.

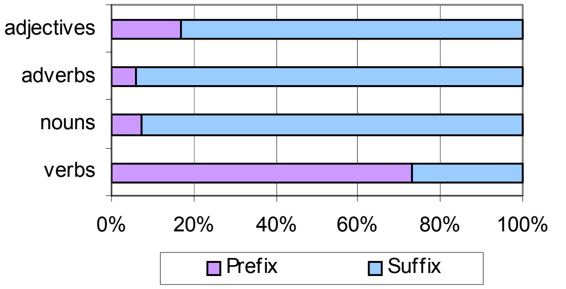

Figure 3: Percentage distribution of words derived by means of prefixes and suffixes.

Table 7: Top 10 most productive affixes to form adjectives, adverbs, nouns, and verbs.

POS	Affixes and counts									
adj.	-ed	-ing	un-	-able	-less	-ive	-y	-ful	-al	in-
	492	226	148	97	80	64	64	50	31	29
adv.	-ly	un-	a-	in-	im-	dis-	-wise	-wards	-	-
	458	14	7	3	2	2	2	1	-	-
noun	-er	-ing	-ness	-or	-ion	-ation	-ment	-ist	-ery	-ity
	607	367	340	79	75	53	45	37	34	32
verb	re-	over-	-en	dis-	un-	de-	out-	mis-	-ize	-ise
	56	34	30	26	22	21	18	18	16	16

3. Conclusions and future work

In this paper we described techniques for finding new sentiment-conveying words, particularly, through synonymy relations and morphologic modifications. Using these methods, it is possible to expand a sentiment lexicon and improve coverage of sentiment analysis systems. In future research we are planning to further increase the SentiFul lexicon by taking into account antonyms (e.g., the reversed scores of '*brave*' could be propagated to its antonyms like '*faint-hearted*', '*caitiff*', '*white-livered*' etc.) and hypernym-hyponym relation (e.g., scores of '*success*' could be propagated to its hyponym '*winning*'). Additionally, compounding using known sentiment-carrying base components might be the efficient way to generate new lexemes (e.g., '*well-wishing*', '*bad-mouth*', '*ill-conditioned*', '*terror-haunted*', etc.). Our primary objective for the future is to implement a procedure for automatically updating the sentiment lexicon.

References

[1] J. Reilly, and L. Seibert. Language and emotion. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (eds.), Handbook of Affective Science, pp. 535–559, 2003.

[2] V. Hatzivassiloglou, and K. R. McKeown. Predicting the semantic orientation of adjectives. Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, pp. 174–181, 1997.

[3] J. Wiebe. Learning subjective adjectives from corpora. Proceedings of the 17th Conference of the AAAI, 2000.

[4] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. Proceedings of 7th Conference on Natural Language Learning, pp. 25–32, 2003.

[5] M. Baroni, and S. Vegnaduzzo. Identifying subjective adjectives through web-based mutual information. Proceedings of the German Conference on NLP, 2004.

[6] K. Moilanen, and S. Pulman. The good, the bad, and the unknown: Morphosyllabic sentiment tagging of unseen words. Proceedings of ACL-08:HLT, pp. 109–112, 2008.

[7] P. D.Turney, and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4):315–346, 2003.

[8] J. Read. Recognising affect in text using pointwise-mutual information. Thesis. University of Sussex, 2004.

[9] S.-M. Kim, and E. Hovy. Determining the sentiment of opinions. Proceedings of Conference on Computational Linguistics, pp. 1367–1373, 2004.

[10] A. Andreevskaia, and S. Bergler. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. Proceedings of the 11th Conference of the European Chapter of the ACL, EACL, 2006.

[11] A. Esuli, and F. Sebastiani. SentiWordNet: a publicly available lexical resource for opinion mining. Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 417–422, 2006.

[12] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Textual affect sensing for sociable and expressive online communication. Proceedings of 2nd International Conference on Affective Computing and Intelligent Interaction, pp. 220–231, 2007.

[13] G. A. Miller. WordNet: An on-line lexical database. International Journal of Lexicography, Special Issue, 3(4):235–312, 1990.

[14] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan, and R. Quirk. Longman Grammar of Spoken and Written English. Pearson Education Limited, 1999.