Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 733 – 743

Procedia
Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew

Yaakov HaCohen-Kerner[a, 1*], Haim Badash[a]

[a]*Dept. of Computer Science, Jerusalem College of Technology, 9116001 Jerusalem, Israel*

Abstract

In this research, given a corpus containing blog posts written in Hebrew and two seed sentiment lists, we analyze the positive and negative sentences included in the corpus, and special groups of words that are associated with the positive and negative seed words. We discovered many new negative words (around half of the top 50 words) but only one positive word. Among the top words that are associated with the positive seed words, we discovered various first-person and third-person pronouns. Intensifiers were found for both the positive and negative seed words. Most of the corpus' sentences are neutral. For the rest, the rate of positive sentences is above 80%. The sentiment scores of the top words that are associated with the positive words are significantly higher than those of the top words that are associated with the negative words.

Our conclusions are as follows. Positive sentences more "refer to" the authors themselves (first-person pronouns and related words) and are also more general, e.g., more related to other people (third-person pronouns), while negative sentences are much more concentrated on negative things and therefore contain many new negative words. Israeli bloggers tend to use intensifiers in order to emphasize or even exaggerate their sentiment opinions (both positive and negative). These bloggers not only write much more positive sentences than negative sentences, but also write much longer positive sentences than negative sentences.

Keywords: Blog corpus; Hebrew; Natural Language Processing; Negative words; Positive words; Seed lists; Sentiment

1. Introduction

The research presented in this paper was performed in the blog domain, which is one of the most popular domains in the Internet. A blog (a truncation of weblog) is a website consisting of informational posts composed by an individual author or a group of authors. The posts are appearing in reverse chronological order (the most recent post appearing first). Blogs typically enable other users to comment or respond to the blog post. Nowadays, there are

[1] Corresponding author. Tel.: +972-2-6751018; fax: +972-2-6751046.
E-mail address: kerner@jct.ac.il

doi:10.1016/j.procs.2016.08.257

hundreds of million public blogs in existence. Processing of blog posts presents challenges due to the large number of words present in the text set, their dependencies and the large number of training documents.

The selected application domain is personal blog posts written in Hebrew. We downloaded a corpus containing blog posts written in Hebrew. Given these blog posts, we are interested to answer the following research questions:

Q1(a). Is it possible to learn new positive words using a basic/extended list of positive words?

Q1(b). Is it possible to learn new negative words using a basic/extended list of negative words?

Q2. Can we discover special groups of words that are associated with the list of positive and negative words?

Q3. What is the distribution of the sentences (neutral, positive, and negative)?

Q4. What are the scores of the top words associated with the positive and negative words and what can we learn from these scores?

To answer these questions, we worked with two seed lists containing sentiment words in Hebrew. These lists were manually generated by us. Each one of these lists contains both positive and negative words. The first list is relatively a small list, containing only 45 words (22 positive and 23 negative). The second list, the largest list, contains 168 words (85 positive and 83 negative). Our motivation to perform experiments with two seed sentiment lists (basic and extended) is to check whether there is any different in the results obtained by these two lists. An example for a question is whether the use of the extended seed sentiment list can discover more positive and negative sentiment words than the use of the basic seed sentiment list.

We defined and activated the following algorithm. Given a blog corpus, we spilt it into sentences. For each sentence, we count the number of positive words (PW) and negative words (NW) included in the sentence according to a given seed sentiment list. Then, we give a sentiment value (+1, -1, 0) to the sentence at hand, according to the value of (PW-NW); i.e., +1 if PW-NW>0, -1 if PW-NW<0, and 0 otherwise. Moreover, for each specific word in the discussed sentence, which is not found in the sentiment list, we add the value of (PW-NW) to the sentiment score of the specific discussed word. After activating this process for all the sentences in the corpus, we have sentiment values for all the words in the corpus, which are not included in the sentiment list. We sorted these words according to their sentiment scores. The words with the highest positive scores are stored in the list of top words associated with positive words, and the words with the lowest negative scores are stored in the list of top words associated with negative words.

This paper is organized as follows: Section 2 supplies relevant background about the Hebrew language, sentiment lexicons, and their expansions, and sentiment blog lexicons and sentiment blog classification. Section 3 presents the two seed sentiment lists that our algorithm works with. Section 4 describes the examined corpus, the experimental results and their analysis. Section 5 presents a summary and proposals for research directions.

2. Relevant background

2.1. The Hebrew language

Hebrew is a Semitic language. It is written from right to left and it uses the Hebrew alphabet. Most Hebrew words are based on three (sometimes four) basic letters, which create the word's stem (root). Except for the word's stem, there are a few other components, which create the word's declensions, such as: belongings, conjugations, objects, prepositions, prefix letters, subjects, terminal letters, and verb types. Overview on these components can be seen at[1].

In Hebrew, it is impossible to find the declensions of a certain stem without an exact morphological analysis based on the components mentioned above.

The English language is richer in its vocabulary than Hebrew. The English language has about 40,000 stems, while Hebrew has only about 3,500 and the number of lexical entries in the English dictionary is 150,000 compared with only 35,000 in the Hebrew dictionary[2].

However, the Hebrew language is richer in its morphology forms. According to linguistic estimates, the Hebrew language has 70,000,000 valid (inflected) forms, while English has only 1,000,000[2]. For instance, the single Hebrew word וכשישתוהו is translated into the following sequence of six English words: "and when they will drink it". In comparison to the Hebrew verb, which undergoes a few changes the English verb stays the same.

In Hebrew, there are up to seven thousand declensions for only one stem, while in English there is only a few declensions. For example, the English word drink has only four declensions (drinks, drinking, drank, and drunk). The relevant Hebrew stem שתה ("drank") has thousands of declensions. Eight of them are presented below: (1) שתיתי

(“I drank”), (2) שתית (“you drank”), (3) שתינו (“we drank”), (4) שותה (“he drinks”), (5) שותים (“they drink”), (6) תשתה (“she will drink”), (7) לשתות (“to drink”), and (8) שתיתיו (“I drink it”).

For more detailed discussions of Hebrew grammar from the viewpoint of computational linguistics, refer to[3]. For Hebrew grammar in Hebrew refer to[1], and in English either to[4] or to[5].

2.2. Sentiment lexicons and their expansions

A sentiment lexicon is a list of positive and negative words and phrases, e.g., “beautiful”, “ugly”, “very good”, “very bad”. Each word or phrase has a positive or negative score reflecting its sentiment polarity. In some cases, a value of +1 represents a positive polarity and a value of -1 represents a negative polarity. In other cases, the value represents not only polarity but also the polarity’s strength. The coverage and the quality of a sentiment lexicon is critical for the success of various tasks, e.g., opinion mining, sentiment analysis, and sentiment classification (Liu, 2012[6]; Feldman, 2013[7]).

Kim and Hovy (2004)[8] automatically identified and estimated sentiments that are combined in opinions. Their system expands two seed lists (positive and negative) by synonyms using WordNet[9,10]. They assume that synonyms (antonyms) of a word have the same (opposite) polarity. The original seed lists contain 44 verbs (23 positive and 21 negative) and 34 adjectives (15 positive and 19 negative). Using synonyms and antonyms for adjectives and only synonyms for verbs, they extracted from WordNet expansions and added them back into the appropriate seed lists. Using these expanded lists, then extracted an additional cycle of verbs and adjectives from WordNet, to obtain finally 12,113 adjectives (5,880 positive and 6,233 negative), and 6,079 verbs (2,840 positive and 3,239 negative).

Automatic estimation of the sentiment score of each word or phrase by current sentiment lexicon learning systems is usually based on propagation methods. These methods typically employ parsing results, syntactic contexts or linguistic information from thesaurus (e.g., WordNet) to calculate the similarity between phrases. For instance, Baccianella et al. (2010)[11] used the glosses information from WordNet, and Velikovich et al. (2010)[12] represented each phrase with its context words derived from web documents.

Qiu et al. (2009)[13] dealt with expansion of a domain sentiment lexicon. They propagate information through both sentiment words and features. Their propagation method exploits the relations between sentiment words and topics or product features that the sentiment words modify, and also sentiment words and product features themselves to extract new sentiment words. The extraction rules are based on relations described in dependency trees. Their experimental results show that their approach is capable to extract many new sentiment words.

Neviarouskaya et al. (2009)[14] presented a system that generates a lexicon for sentiment analysis. The authors described methods that automatically generate and score a new sentiment lexicon, called SentiFul, and expand it through direct synonymy relations and morphologic modifications with known lexical units.

Liu et al. (2011)[15] suggested a method to build Chinese sentiment lexicon using HowNet[16]. “HowNet is an on-line common-sense knowledgebase unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons of the Chinese and their English equivalents”[7]. Using Chinese basic sentiment words, a corpus, and HowNet, they can identify sentiment words and expand their sentiment lexicon. Their method is based on analysis of sentence structure and calculations of semantic similarity scores. A Chinese text sentiment orientation classification experiment using this lexicon obtained above 70% accuracy.

Lu et al. (2011)[17] automatically generated a context-dependent sentiment lexicon from unlabeled opinionated text documents. Their method can learn new domain specific sentiment words and aspect-dependent sentiment. For a given domain, their system can improve the coverage of a general sentiment lexicon and performance of sentiment classification can be significantly improved with the automatically generated context-dependent sentiment lexicon.

Tang et al. (2014)[18] described the construction of a large-scale twitter-specific sentiment lexicon. Their method is composed of two components: (1) a representation learning algorithm that learns the embedding of phrases, which are used as features for classification and (2) a seed expansion algorithm that enlarges a small list of sentiment seeds to obtain training data for constructing the phrase-level sentiment classifier.

2.3. Sentiment blog lexicons and sentiment blog classification

Chesley et al. (2006)[19] used verbs and adjectives and a classifier they developed to classify sentiment blog posts. They used (1) an automatic text analyzer, called Semantex (Srihari et al. 2006[20]) that groups verbs according to classes that often correspond to their polarity classification, and (2) Wiktionary[21], the Wikipedia's online dictionary, to determine the polarity of adjectives extracted from the blog posts.

Godbole et al. (2007)[22] presented a system that assigns scores indicating positive or negative opinion to each entity in the text corpus. Their system consists of a sentiment identification phase, which associates expressed opinions with each relevant entity, and a sentiment aggregation and scoring phase, which scores each entity relative to others in the same class. Finally, they evaluated the significance of their scoring techniques over a large corpus of news and blogs.

Melville et al. (2009)[23] used background lexical information in terms of word-class associations, and refine this information for specific domains using any available training examples. They incorporated the lexical knowledge in supervised learning for blog classification. Empirical results on various areas show that their method performs better than using only background knowledge or only training data.

3. The seed sentiment lists

As mentioned above, we prepared two lists containing sentiment words in Hebrew. Each one of them contains both positive and negative words. The first list is a basic list containing only 45 words (22 positive and 23 negative). The second list contains 168 words (85 positive and 83 negative). Table 1 presents the basic sentiment list and Table 2 presents the extended sentiment list. Each entry in the table includes the index number of the sentiment word, the word in Hebrew, and its translation into English.

Table 1. The basic (small) sentiment list.

Positive words						Negative words					
#	Hebrew	English	#	Hebrew	English	#	Hebrew	English	#	Hebrew	English
1	התלהבות	enthusiasm	12	כיף	fun	1	דאגה	concern	13	פסימי	pessimistic
2	מצוין	excellent	13	טוב	good	2	היסוס	hesitation	14	עצוב	sad
3	אופטימי	optimistic	14	שמח	happy	3	עצב	sadness	15	איום	terrible\threat
4	מעולה	superior	15	תקווה	hope	4	רע	bad	16	אסון	disaster
5	נפלא	wonderful	16	מאושר	joyful	5	מתלונן	complainant	17	ייאוש	despair
6	יתרון	advantage	17	נהדר	magnificent	6	גינה	criticize	18	דיכאון	depression
7	מותר	allowed	18	שיבח	praised	7	ייאוש	desperation	19	סבל	suffering
8	מהמם	amazing	19	בטחון	safety	8	מדוכדך	despondent	20	זוועה	horror
9	אושר	bliss	20	הצלחה	success	9	חיסרון	disadvantage	21	אסור	forbidden
10	ביטחון	confidence	21	ניצח	win	10	כישלון	failure	22	גרוע	inferior
11	נהנה	enjoy	22	שמחה	happiness	11	כשל	failure	23	סבל	suffer
						12	הפסיד	lose			

Table 2. The extended (large) sentiment list.

Positive words						Negative words					
#	Hebrew	English	#	Hebrew	English	#	Hebrew	English	#	Hebrew	English
1	שלם	complete	44	עניו	humble	1	חושש	afraid	44	מעציב	saddening
2	מצוין	excellent	45	שיפור	improvement	2	רע	bad	45	מרפה	slacks
3	חבל על הזמן	fantastic	46	מרבה	increase	3	היסוס	hesitation	46	מתנשא	snobbish
4	אנרגיות חיוביות	positive energy	47	מבין עניין	know what is going on	4	יש מקום לשיפור	there is room for improvement	47	קלקול	spoilage
5	ממגנט	magnetize	48	חביב	likeable	5	עצב	sadness	48	מקלקל	spoils
6	אופטימי	optimistic	49	אהבה	love	6	מפחית	subtract	49	חשוד	suspicious
7	עוצמתי	powerful	50	אוהב	lover	7	שגוי	wrong	50	כעור	ugly

8	חזק	strong	51	נאמן	loyal	8	אפאטיות	apathy	51	מכוער	ugly
9	מעולה	superior	52	נאה	nice	9	שבור	broken	52	מבאס	upset
10	נצחון	victory	53	תקין	normal	10	לא יכול	can't	53	איום	warning
11	מוסיף	add	54	שלווה	peaceful	11	עגמימות	cheerless	54	רפה	weak
12	יתרון	advantage	55	נעים	pleasant	12	תלונה	complaint	55	נסוג	withdraw
13	מותר	allowed	56	עממי	popular	13	מגנה	criticize	56	נסיגה	withdrawal
14	מהמם	amazing	57	משמח	gladdening	14	גינה	criticized	57	מתנדנד	wobbles
15	יפה	beautiful	58	משבח	praise	15	עקום	crook	58	חשש	worry
16	יופי	beauty	59	שיבח	praised	16	קללה	curse	59	שוגה	wrong
17	אהוב	beloved	60	מעלה	raise	17	ייאוש	despair	60	כועס	angry
18	מבורך	blessed	61	מתקן	reformer	18	מדוכדך	despondent	61	יהיר	arrogant
19	בירך	blessed	62	רגוע	relaxed	19	חיסרון	disadvantage	62	מקולל	cursed
20	מאושר	blissful	63	כיבוד	respecting	20	דאון	down	63	תבוסה	defeat
21	בטוח	certain	64	צודק	right	21	מוריד	download	64	דחוי	deferred
22	עליז	cheerful	65	יציב	stable	22	להיכשל	fail	65	ייאוש	despair
23	עליזות	cheerfulness	66	ישר	straight	23	כישלון	failure	66	הרס	destruction
24	פיקח	clever	67	מחזק	strengthen	24	פחד	fear	67	אכזבה	disappointment
25	הודאה	confession	68	הצלחה	success	25	סכל	fool	68	מגעיל	disgusting
26	ביטחון	confidence	69	מוצלח	successful	26	שנאה	hate	69	זלזול	disrespect
27	התחשבות	consideration	70	טעים	tasty	27	שונא	hater	70	אגואיסט	egoist
28	בנייה	construction	71	מתחשב	thoughtful	28	קילל	to curse	71	כישלון	failure
29	מתוקן	corrected	72	להצליח	to succeed	29	חסר	lack	72	אסור	forbidden
30	תיקון	correction	73	אמת	truth	30	עצלן	lazy	73	שנוא	hated
31	חרוץ	diligent	74	טוב מאוד	very good	31	ממעיט	lessen	74	בלתי נסבל	insufferable
32	סוף העולם	end of the road	75	יחסית טוב	relatively well	32	אנרגיות שליליות	negative energies	75	בלתי אפשרי	impossible
33	נהנה	enjoy	76	חלש	weak	33	הפסד	loss	76	סבל	suffering
34	מגדיל	enlarge	77	ניצח	Win	34	הפסיד	lost	77	נורא ואיום	terrible
35	אמון	faith	78	ניצחון	Victory	35	שקר	lie	78	גרוע	lousy
36	כיף	fun	79	יכול	Can	36	לא טוב	not good	79	בגידה	treason
37	כיפי	funny	80	טוב	Good	37	מכשול	obstacle	80	כיעור	ugliness
38	השתפר	get better	81	שמחה	happiness	38	מקולקל	out of order	81	על-הפנים	very bad
39	משמח	gladdening	82	אפשר	Possible	39	פסימי	pessimistic	82	לא כדאי	inexpedient
40	ברכה	greeting	83	כדאי	worthwhile	40	מקטין	reduce	83	לא נעים	unpleasant
41	שמח	happy	84	נכון	Right	41	דוחה	repulsive			
42	תקווה	hope	85	מקפצה	springboard	42	עגום	rueful			
43	קיווה	hoped				43	עצוב	sad			

4. The examined corpus and experimental results

We downloaded a corpus containing blog posts written in Hebrew from http://israblog.nana10.co.il/. This blog corpus contains 100,514 documents, 11,406,047 sentences, and 50,515,843 words. Sub-sections 4.1 and 4.2 introduce the experimental results for the blog corpus using the small and the large sentiment lists, respectively.

4.1. Experimental results for the blog corpus using the small sentiment list

In Tables 3 and 4, we present the top 50 words that are associated with positive words and negative words included in the small sentiment list (Table 1), respectively.

Table 3. Top words that are associated with the small list of positive words.

#	Hebrew	English	Score	#	Hebrew	English	Score
1	לי	for me	47249	26	אם	if	7126
2	אני	me	45924	27	כמה	a few	6862
3	אתה (את)	you	37334	28	שיהיה	that will be	6549
4	זה	this	35197	29	כבר	already	6396
5	היה	was	27990	30	עוד	more	6330
6	יותר	more	19957	31	הזה	this	6135
7	אבל	but	19054	32	כמו	like	5860
8	כל	all	18378	33	מזל	luck	5786
9	עם	with	18332	34	איזה	which	5751
10	על	on	18030	35	קצת	a little	5714
11	שלי	mine	17289	36	לו	to him	5598
12	אז	then	14596	37	שזה	that this	5593
13	רק	only	14159	38	אחד	one	5437
14	מה	what	12978	39	לך	for you	5378
15	הוא	he	11435	40	הכל	everything	5315
16	כי	because	11392	41	משהו	something	5178
17	הכי	the most	10470	42	עכשיו	now	5240
18	כך	like this	10119	43	מאוד	very	5199
19	יש	there is	8747	44	אנשים	people	5153
20	יום	day	8550	45	רוצה	want	5160
21	אותי	me	8039	46	שהוא	that he	5117
22	הרבה	a lot	7955	47	באמת	indeed	4978
23	להיות	to exist	7927	48	אותו	him	4937
24	יהיה	will be	7772	49	שם	there	4904
25	היום	today	7363	50	וזה	and this	4830

Table 4. Top words that are associated with the small list of negative words.

#	Hebrew	English	Score	#	Hebrew	English	Score
1	נורא	tribble	-145	26	מתמשך	ongoing	-11
2	אסור	forbidden	-85	27	וייסורים	and suffering	-11
3	הכי	most	-47	28	שפגעתי	that I harmed	-10
4	אל	not	-34	29	בדידות	loneliness	-10
5	ומר	and Mr., and bitter	-30	30	להתייאש	to despair	-10
6	בכי	crying	-27	31	הסירו	they remove	-10
7	חרדה	anxiety	-25	32	נובל	wither	-9
8	קליני	clinical	-19	33	מוטל	imposed	-9
9	טבלת	table	-18	34	פגוע	damaged	-9
10	חרוץ	diligent	-18	35	ובודד	and lonely	-9
11	ואסור	and forbidden	-17	36	ודואג	and worried	-9
12	ויותר	and more	-17	37	ואכזבה	and disappointment	-9
13	האיסור	the prohibition	-17	38	ולשנוא	and to hate	-8
14	פוחד	afraid	-15	39	מוחץ	crushing	-8

15	קיומי	existential	-15	40	שקיעת	sunset of	-8
16	בתכלית	in purpose	-15	41	אסירים	prisoner	-8
17	לקטוף	pick to	-14	42	כשהלב	when the heart	-8
18	ובדידות	and loneliness	-14	43	ותסכול	and frustration	-8
19	התאבדות	suicide	-13	44	יאונה	occurred	-8
20	התאומים	the twins	-13	45	סיפור	story	-7
21	ומדוכא	and depressed	-13	46	שקשה	that difficult	-7
22	ומגעיל	and disgusting	-12	47	לזלזל	to disparage	-7
23	כשהמלאכים	when the angels	-12	48	להתעייף	to get tired	-7
24	ענוג	delicate	-12	49	מהדהד	echo	-7
25	בכו	they cried	-11	50	ופתטי	and pathetic	-7

To answer the research questions presented in the introduction section, we analyze various statistics including the results that are introduced in Tables 3 and 4, which are based on the short sentiment list (Table 1).

A1 (Answer to Q1). Only one new positive word (מזל, luck, #33) has been discovered in Table 3. However, according to Table 4, 24 new negative words (almost half of the 50 top words!) have been discovered: (נורא, tribble, #1), (אסור, forbidden, #2), (בכי, crying, #6), (חרדה, anxiety, #7), (ואסור, and forbidden, #11), (האיסור, the prohibition, #13), (פוחד, afraid, #14), (ובדידות, and loneliness, #18), (התאבדות, suicide, #19), (נורא, and depressed, #21), (ומגעיל, and disgusting, #22), (בכו, they cried, #25), (וייסורים, and suffering, #27), (שפגעתי, that I harmed, #28), (בדידות, loneliness, #29), (להתייאש, to despair, #30), (פגוע, damaged, #34), (ובודד, and lonely, #35), (ודואג, and worried, #36), (ואכזבה, and disappointment, #37), (ולשנוא, and to hate, #38), (ותסכול, and frustration, #43), (שקשה, that difficult, #46), and (לזלזל, to disparage, #47).

A2. Analysis of the 50 top words (Table 3) that were obtained using the positive seed words, leads to the discovery of a few special groups of words. The first group contains four words that are first-person pronoun(s) and words that are relevant to pronoun(s),that have been discovered in relatively high ranks: (לי, me, #1), (אני, I, #2), (שלי, mine, #11), and (אותי, a term used to indicate a direct object, #21). The second group contains third-person pronoun and words that are relevant to these pronoun: (היה, was, #5), (הוא, he, #15), (יהיה, will be, #24), (לו, to him, #36), (שהוא, that he, #46), and (אותו, him, #48). A third special group contains 6 intensifiers: (יותר, more, #6), (הכי, the most, #17), (כמה, a few, #27), (עוד, more, #30), (הכל, everything, #40), and (מאוד, very, #43).

Analysis of the 50 top words (Table 4) that were obtained using the negative seed words, did not find any pronouns and related words that are relevant to pronouns. We did find a special group contains 5 intensifiers: (הכי, the most, #3), (ויותר, and more, #12), (מתמשך, ongoing, #26), (מוחץ, crushing, #39), and (שקשה, that difficult, #46).

Answers A1 and A2 point that positive sentences more "refer to" the authors themselves (first-person pronouns and related words) and are also more general, e.g., more related to other people (third-person pronouns), while negative sentences are much more concentrated on negative things and therefore contain many new negative words. The Israeli bloggers tend to use intensifiers in their sentiment sentences to emphasize or even exaggerate their sentiment opinions (both positive and negative).

A3. We discovered that most of the sentences are neutral (around 97.8%). There 229,961 positive sentences (around 2%) and only 48,074 negative sentences (around 0.42%). There are 4.7 times more positive sentences than negative sentences. A possible explanation to this finding is that Israeli bloggers prefer to write much more about positive things than negative things, especially when it comes to their personal blog posts that are publicly available.

A4. The scores (in absolute values) of the 50 top words (Table 3) that are associated with the small list of positive words are significantly higher than the scores of the 50 top words (Table 4) that are associated with the small list of negative words. One main reason for this finding is that the number of positive sentences is 4.7 times more than the number of negative sentences. The score of the first four words that are associated with the positive words is higher than 35,000, while the score of the first four words that are associated with the negative words is only lower than -33. The score of the last word (ranked at place #50) in Table 3 is 4,830, while the score of the last word in Table 4 is only -7. An additional explanation to this finding might be that an average positive sentence includes much more words than an average negative sentence. That is to say, Israeli bloggers not only write much more about positive things than negative things, but also write much longer positive sentences than negative sentences from the viewpoint of number of words.

4.2. Experimental results for the blog corpus using the large sentiment list

In Tables 5 and 6, we present the top 50 words that are associated with positive words and negative words included in the large sentiment list (Table 2), respectively.

Table 5. Top words that are associated with the large list of positive words.

#	Hebrew	English	Score	#	Hebrew	English	Score
1	לא	no	114222	26	יש	there is	18663
2	את	you	92641	27	רק	only	17246
3	אני	I	92020	28	היא	she	17241
4	לי	me	80915	29	כמה	many, how much	16285
5	זה	this	75037	30	שלא	that not	14836
6	של	of	46113	31	שהוא	that he	14502
7	היה	was	45140	32	הרבה	many	14438
8	אבל	but	42539	33	יום	day	14279
9	על	on	42437	34	כבר	already	14159
10	כל	all	39193	35	כמו	like	14041
11	יותר	more	37958	36	אחד	one	13968
12	עם	with	35971	37	שזה	that this	13461
13	שלי	mine	35115	38	עוד	more	13227
14	מה	what	30590	39	אותו	him	12975
15	הוא	he	28667	40	הזה	this	12938
16	ממש	really	27932	41	לו	him	12861
17	אז	so	26941	42	מאוד	very	12740
18	כי	because	24316	43	היום	today	12120
19	להיות	to be	22905	44	רוצה	want	12055
20	גם	also	21479	45	יהיה	will be	11549
21	כך	like this	20955	46	משהו	something	11396
22	אם	if	20211	47	לעשות	to do	11221
23	אותי	me	19786	48	לך	go, to you	11152
24	הכי	most	19246	49	שם	there	11029
25	או	or	19240	50	באמת	really	10632

Table 6. Top words that are associated with the large list of negative words.

#	Hebrew	English	Score	#	Hebrew	English	Score
1	הפנים	interior	-1797	26	שונא	hate	-18
2	בלתי	non	-807	27	תהומית	abysmal	-15
3	אפשרי	possible	-705	28	עצבנות	nervousness	-15
4	נסבל	sufferable	-414	29	וחורבן	and destruction	-15
5	תועלת	benefit	-155	30	וחרדה	and anxiety	-14
6	ואיום	and terrible	-153	31	וכזב	and lie	-14
7	פואנטה	intention	-104	32	דחפה	she pushed	-14
8	אונים	strength	-89	33	וזלגה	and she trickled	-14
9	משמעות	meaning	-79	34	נטישה	abandonment	-13

10	תכלית	purpose	-79	35	תקליט	album	-13
11	ונורא	and horrible	-66	36	האיסור	the prohibition	-12
12	קיבה	stomach	-44	37	מהחושך	from the dark	-11
13	גבהים	heights	-36	38	איכס	disgust	-11
14	ודוחה	and repulsive	-35	39	רוגז	anger	-11
15	תקדים	precedent	-32	40	כשהמלאכים	when the angel	-11
16	בושה	shame	-31	41	ההבעה	the expression	-11
17	ומגעיל	and disgusting	-29	42	ממחטים	needle	-11
18	תסמונת	syndrome	-27	43	פחד	fear	-10
19	והתחמקת	and you evaded	-23	44	תפסו	caught	-10
20	מרוח	spread	-21	45	ומסריח	and smelly	-10
21	ומר	and Mr., and bitter	-21	46	מהערך	from the value	-10
22	מדבר	desert	-21	47	עסיסית	juicy	-10
23	ישע	salvation	-20	48	רפה	weak	-10
24	טקט	tact	-19	49	הר	mountain	-9
25	שמפחיד	that scares	-18	50	ואלון	and oak	-9

The answers to the research questions presented in the introduction section, related to the large sentiment list (Table 2), based on the results that are introduced in Tables 5 and 6 are as follows.

A1 (Answer to Q1). No positive word was discovered among the top words that are associated with the large list of positive words in Table 5. However, according to Table 6, 25 new negative words (half of the 50 top words!) have been discovered: (בלתי, non, #2), (נסבל, sufferable, #4), (ואיום, and terrible, #6), (ודוחה, and repulsive, #14), (בושה, shame, #16), (ומגעיל, and disgusting, #17), (תסמונת, syndrome, #18), (והתחמקת, and you evaded, #19), (שמפחיד, that scares, #25), (שונא, hate, #26), (תהומית, abysmal, #27), (עצבנות, nervousness, #28), (וחורבן, and destruction, #29), (וחרדה, and anxiety, #30), (וכזב, and lie, #31), (דחפה, she pushed, #32), (נטישה, abandonment, #34), (האיסור, the prohibition, #36), (מהחושך, from the dark, #37), (איכס, disgust, #38), (רוגז, anger, #39), (פחד, fear, #43), (תפסו, caught, #44), (ומסריח, and smelly, #45)), and (רפה, weak, #48).

The results obtained by the two sentiment lists were very similar. Almost no new positive words were discovered by these two lists (one new positive word and zero new positive words in Tables 3 and 5, respectively) among the top 50 words are associated with the basic/extended list of positive words. In contrast, about half of the top 50 words (24 new negative words and 25 new negative words in Tables 4 and 6, respectively) that were associated with the basic/extended list of negative words were discovered as new negative words.

A2. Analysis of the 50 top words (Table 5) that were obtained using the positive seed words, leads to the discovery of a few special groups of words. The first group contains four words, one first-person pronoun and words that are relevant to this pronoun: (אני, I, #3), (לי, me, #4), (שלי, mine, #13), and (אותי, a term used to indicate a direct object, #23). The second group contains third-person pronouns and words that are relevant to these pronouns: (הוא, he, #15), (היא, he, #28), (שהוא, that he, #31), (אותו, him, #39) and (לו, to him, #41). A third special group contains 6 intensifiers: (כל, all, #10), (יותר, more, #11), (גם, also, #20), (הכי, the most, #24), (כמה, a few, #29), (הרבה, many, #32), (עוד, more, #38), (מאוד, very, #42), and (באמת, really, #50).

Analysis of the 50 top words (Table 6) that were obtained using the negative seed words, did not find any pronouns and related words, but did find 3 intensifiers: (גבהים, heights, 14), (תהומית, abysmal, #28), and (עסיסית, juicy, #46).

Also in this experiment, answers A1 and A2 point that positive sentences more "refer to" the authors themselves (first-person pronouns and related words) and are also more general, e.g., more related to other people (third-person pronouns), while negative sentences are much more concentrated on negative things and therefore contain many new negative words. The Israeli bloggers tend to use intensifiers in their sentiment sentences to emphasize or even exaggerate their sentiment opinions (both positive and negative).

A3. Most of the sentences are neutral (around 95.43%). There are 425,262 positive sentences (around 3.7%) and only 99,717negative sentences (around 0.87%). There are 4.2 times more positive sentences than negative sentences (comparing to 4.7 in Sub-Section 4.1 regarding the results based on the small sentiment list). Also here, a possible explanation to this finding is that Israeli bloggers prefer to write much more about positive things than negative things, especially when it comes to their personal blog posts that are publicly available.

A4. The scores (in absolute values) of the 50 top words (Table 5) that are associated with the large list of positive words are significantly higher than the scores of the 50 top words (Table 6) that are associated with the large list of negative words. One main reason for this finding is that the number of positive sentences is 4.2 times more than the number of negative sentences. The score of the first five words that are associated with the positive words is higher than 75,000, while the score of the first five words that are associated with the negative words is only lower than -154. Again, an additional explanation to this finding might be that an average positive sentence includes much more words than an average negative sentence. That is to say, Israeli bloggers not only write much more about positive things than negative things, but also write much longer positive sentences than negative sentences from the viewpoint of number of words.

5. Summary and future work

We presented a working system that analyzed a blog corpus written in Hebrew from the viewpoint of its positive and negative sentiment words. The answers to the research questions mentioned in Section 1, based on the results of both experiments (small and large sentiment lists) were very similar as follows: We discovered many new negative words (around half of the top 50 words) but only one positive word. The new discovered negative words and one new positive word can enrich the seed sentiment lists and by that improve future sentiment analysis systems as well as other linguistic applications for the Hebrew language.

Among the top words that are associated with the positive seed words, we discovered various first-person and third-person pronouns. Intensifiers were found for both the positive and negative seed words. Most of the corpus' sentences are neutral. For the rest, the rate of positive sentences is above 80%. The sentiment scores of the top words that are associated with the positive words are significantly higher than those of the top words that are associated with the negative words. The special groups of words that have been discovered (first-person and third-person pronouns, and intensifiers) might help in future studies and systems to recognize new positive and negative words in their environment.

Our conclusions about the tested blogs are as follows. Positive sentences more "refer to" the authors themselves (first-person pronouns and related words) and are also more general, e.g., more related to other people (third-person pronouns), while negative sentences are much more concentrated on negative things and therefore contain many new negative words. Israeli bloggers tend to use intensifiers in order to emphasize or even exaggerate their sentiment opinions (both positive and negative). Finally, these bloggers not only write much more positive sentences than negative sentences, but also write much longer positive sentences than negative sentences.

Possible directions for future research are: (1) defining improved sentiment lists from the following viewpoints: giving sentiment scores for each word, associating a suitable PoS (Part of Speech) tag with each word, words' normalization in the sense of removal of affixes; removal of prefix letters, single/many, male/female, dealing with abbreviations as done in various studies[24,25,26], etc.; (2) conducting additional experiments using much larger blog posts in Hebrew; (3) extending the experiments to other languages and to see what are the similarities and differences between the Israeli bloggers and bloggers from other countries who write in other languages; and (4) extending the experiments to news corpora written in Hebrew and other languages.

Acknowledgements

The authors would like to thank Yiska HaCohen-Kerner, Menahem darmoni, and Yehuda Goldstoph for their support. We would also like to thank the two anonymous reviewers for their useful and instructive comments.

References

[1] Yelin, D.: Dikduk HaLason HaIvrit (Hebrew Grammar, in Hebrew), Jerusalem; 1970.

2 Choueka, Y., Conley E.S., Dagan I.: A Comprehensive Bilingual Word Alignment System: Application to Disparate Languages – Hebrew and English, in J. Veronis (Ed.), Parallel Text Processing, Kluwer Academic Publishers, 2000, pp. 69-96.

3 Wintner, S.: Hebrew Computational Linguistics: Past and Future, Artificial Intelligence Review, 2004, 21(2): 113-138.

4 Glinert, L.: Hebrew – An Essential Grammar, Routledge, London; 1994.

5 Wartski, I.: Hebrew Grammar and Explanatory Notes, The Linguaphone Institute, London; 1900.

[6] Liu, Bing. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 2012, 5(1):1–167.

[7] Feldman, R. Techniques and applications for sentiment analysis. Communications of the ACM, 2013, 56(4): 82-89.

[8] Kim, S. M., Hovy, E. Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics; 2004, p. 1367.

[9] Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An On-Line Lexical Database. http://www.cosgi.princeton.edu/~wn. 1993.

[10] Fellbaum, C., D. Gross, and K. Miller. Adjectives in WordNet. http://www.cosgi. princeton.edu/~wn. 1993.

[11] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC, 2010, 10: 2200–2204.

[12] Velikovich, Leonid, Blair-Goldensohn, Sasha, Hannan, Kerry, McDonald, Ryan. The viability of web derived polarity lexicons. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 777–785.

[13] Qiu, G., Liu, B., Bu, J., Chen, C. Expanding Domain Sentiment Lexicon through Double Propagation. In *IJCAI*, 2009, 9: 1199-1204.

[14] Neviarouskaya, A., Prendinger, H., Ishizuka, M. Sentiful: Generating a reliable lexicon for sentiment analysis. In Affective Computing and Intelligent Interaction and Workshops, ACII 2009. 3rd International Conference on IEEE, 2009, pp. 1-6.

15 Liu, Y. W., Xiao, S. B., Wang, T., Shi, S. C. Building Chinese sentiment lexicon based on HowNet. In Advanced Materials Research, 2011, 187: 405-410.

[16] Dong, Z., Dong, Q. HowNet and the Computation of Meaning . Singapore: World Scientific, 2006, pp. 85-95.

[17] Lu, Y., Castellanos, M., Dayal, U., Zhai, C. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In Proceedings of the 20th international conference on World wide web, ACM; 2011, pp. 347-356.

[18] Tang, D., Wei, F., Qin, B., Zhou, M., Liu, T. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. In COLING; 2014, pp. 172-182.

[19] Chesley, P., Vincent, B., Xu, L., & Srihari, R. K. Using verbs and adjectives to automatically classify blog sentiment. Training, 2006, 580(263), 233.

[20] Srihari, R.; Li, W.; Niu, C.; and Cornell, T. InfoXtract: A Customizable Intermediate Level Information Extraction Engine. Journal of Natural Language Engineering; 2006, 12:4.

[21] http://en.wiktionary.org/wiki/ .

[22] Godbole, N., Srinivasaiah, M., and Skiena, S. Large-Scale Sentiment Analysis for News and Blogs. ICWSM; 2007, 7(21): 219-222.

[23] Melville, P., Gryc, W., and Lawrence, R. D. Sentiment analysis of blogs by combining lexical knowledge with text classification
In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM; 2009. pp. 1275-1284.

[24] HaCohen-Kerner, Y., Kass, A., and Peretz, A. Combined one sense disambiguation of abbreviations. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers Association for Computational Linguistics., 2008, pp. 61-64.

[25] HaCohen-Kerner, Y., Kass, A., and Peretz, A. HAADS: A Hebrew Aramaic abbreviation disambiguation system. Journal of the American Society for Information Science and Technology; 2010, 61(9): 1923-1932.

[26] HaCohen-Kerner, Y., Kass, A., and Peretz, A. Initialism disambiguation: Man versus machine. Journal of the American Society for Information Science and Technology, 2013; 64(10), 2133-2148.