Dictionary Based Amharic Sentiment Lexicon Generation

Girma Neshir Alemneh[1(✉)], Andreas Rauber[2], and Solomon Atnafu[3]

[1] IT Doctoral Program, Addis Ababa University, Addis Ababa, Ethiopia
girma1978@gmail.com
[2] Institute of Information Systems Engineering,
Technical University of Vienna, Vienna, Austria
rauber@ifs.tuwien.ac.at
[3] Department of Computer Science,
Addis Ababa University, Addis Ababa, Ethiopia
solomon.atnafu@aau.edu.et

Abstract. Sentiment analysis is a hot research area with several applications including analysis of political opinions, classifying comments, movie reviews, news reviews and product reviews. To employ rule based sentiment analysis, sentiment lexicon is required. However, manual construction of a sentiment lexicon is time consuming and costly for resource-limited languages. To reduce development time and costs, we propose an algorithm for constructing Amharic sentiment lexicons. The proposed approach transfers sentiment labels from a one language (e.g. English) to resource-limited language (e.g. Amharic) relying on Amharic-English dictionary. Using Bilingual/Monolingual dictionaries as a bridge, two Amharic sentiment lexicons are automatically generated the first based on SO-CAL polarity lexicon, the second on SentiWordNet 3.0. For each Amharic word, the algorithm finds the meaning of the corresponding English word(s). For these English words, sentiment information is searched from the aforementioned sentiment lexicon(s). The weighted average of returned sentiment values, part of speech and gloss information is assigned to the Amharic word. Lexicons of 5683 and 13679 words, respectively, are generated automatically and evaluated subsequently.

Keywords: Lexicon generation algorithm · Sentiment analysis · Amharic sentiment lexicon · Sentiment lexicon generation

1 Introduction

Sentiment Analysis or opinion mining is the process of detecting subjective or emotional indicators in the text. That is finding the attitude of the author towards a particular topic, event, entity, aspect or issue. Nowadays, due to the advancement of web technology, users are not only consumers of online information but also they are producer of information. Most of these data carry opinions that need further analysis to detect its polarity levels (i.e. negative and positive). For instance, political parties want to gather opinions of voters, or a government wants to collect summarized opinions for improvement towards their new policy or existing public services. Thus, a sentiment lexicon is a

F. Mekuria et al. (Eds.): ICT4DA 2019, CCIS 1026, pp. 311–326, 2019.
https://doi.org/10.1007/978-3-030-26630-1_27

valuable resource that should be readily available for opinion mining tasks. So far, sentiment analysis research works were restricted to dominant languages (mostly English). Currently, opinionated language resources are increasingly generated and used in languages other than English. Amharic is one of these resource-limited languages that has started to get attention. But, still it has no sufficient linguistic resources for sentiment analysis such as part of speech taggers and sentiment lexicons. A sentiment lexicon contains a list of opinion words where each term has been assigned positive and negative values, and sometimes, part of speech and glosses definition are included. This work aims to address the following research questions: (a) how do we develop Amharic sentiment lexicons by employing monolingual and cross-lingual resources automatically? (b) how do we evaluate and validate the quality and accuracy of Amharic polarity lexicon's subjectivity clues? (c) how can we measure the accuracy and the coverage of the generated Amharic Sentiment Lexicons for detecting the subjectivity of Amharic text?

The remaining part of the article is organized as follows: in Sect. 2, we provide an overview of related work. Section 3 describes the data sets and the proposed methods. In Sect. 4, we present results and discussions, followed by a summary of our conclusions in Sect. 5.

2 Related Work

This section briefly presents a few key related works. Gebremeskel [6] manually built Amharic polarity Lexicon of 900 terms. This was, again manually, extended by Tilahun to 1,000 terms [7]. In these works, the polarity value of each term in this Amharic sentiment lexicon was labeled either negative 2 or positive 2 without using numerical value in between or no nominal fine-grained levels of sentiment strength granularities or no part of speech information in the lexicon. The authors used this lexicon for lexicon based (rule based) Amharic opinion mining tasks. However, the effort required to further extend these lexicons manually is prohibitive, calling for automated methods, benefiting from the efforts invested in building such dictionaries in other languages. Taboada et al. [1] propose a semantic orientations calculator (SO-CAL) for lexicon based subjectivity classification. SO-CAL utilizes sentiment dictionaries and also includes intensification and negation to calculate and assign semantic orientation of words in the text. The major strength of this work is that the lexicon performed consistently in any domain to extract word level/sentence level opinion in text and also presented the process of creating and evaluating sentiment dictionaries using Mechanical Turk. SO-CAL performed well in sentiment analysis of user's blog postings. We will use SO-CAL as one of the baseline sentiment lexicons for assigning sentiment values to our translated terms.

Medagoda et al. in [2] built SentiWordNet 3.0 for Sinhala language by mapping English SentiWordNet 3.0 relying on the online English-Sinhala dictionary. The English words in the dictionary were used as search key to generate a lexicon for the corresponding translated Sinhala sentiment lexicon. If the translated Sinhala word is found, then it is inserted with its polarity value and POS tag into the Sinhala sentiment lexicon, otherwise search proceeds for the next English word in the dictionary. The final Sinhala Sentiment Lexicon contains 5973 adjectives and 405 adverbs. This lexicon was evaluated using 2,083 manually classified news article opinions collected from Sinhala online newspaper. Based on the different classification methods using this lexicon, the accuracy achieved was between 56–60%. As this evaluation result is below the baseline, handling negations, multiword with negations and context sensitivity were suggested to be addressed in the forthcoming works to improve the performance of the lexicon generation. We will use a similar approach to generate Amharic Sentiment Lexicon, considering the idiosyncrasies of the Amharic language.

3 Methods

3.1 Data Sets and Lexical Resources

This subsection describes the main data sets and lexical resources used for building and evaluating Amharic Sentiment Lexicons.

(a) English SentiWordNet 3.0: SentiWordNet is automatically built based on synsets of wordnet version 3.0 [3]. This lexicon contains 72,092 terms with part of speech, id number, PosScore, NegScore, SynsetTerms and Gloss. The pair (POS, ID) uniquely identifies a WordNet (3.0) synset. The values PosScore and NegScore are the positivity and negativity scores (in the range of 0 to1) assigned to each entry of the synset. The objectivity score is computed as: ObjScore = 1 − (PosScore + NegScore) SynsetTerms column reports the terms, with sense number, belonging to the synset (separated by spaces). We selected to port this lexicon into Amharic because of its extensive coverage.

(b) SO-CAL Polarity Lexicon: SO-CAL refers to the Semantic Orientation CALculator, a tool to extract sentiment from text. It has a long history of development. We use this lexicon [1] as a baseline. SO-CAL contains 10126 words with polarity value ranges from −5 to +5. The lexicon is categorized into adjectives, Adverbs, nouns, verbs and intensifier word lists. This lexicon has been extensively tested showing good performance in different domains.

(c) Amharic-English Dictionary: This is a dictionary that works in one direction (i.e. Amharic to English). For each Amharic word, it contains part-of-Speech tag(s) and corresponding meanings in English. The English text might be a single word, phrase, or list of synonyms. The total size of this dictionary is more than 31000 terms which are obtained by merging Amharic-English dictionary (12700 Amharic words) [4], Amsalu_Aklilu Amharic-English Dictionary (16231 words) and Amharic-English dictionary by SelamSoft Plc (1,075 words). This dictionary serves as a bridge to propagate sentiment from English sentiment lexicon to Amharic sentiment lexicon.

(d) Amharic-Amharic Dictionary: From more than 30 thousand entries of Amharic lexical wordlist in [4], we built 33965 Amharic-Amharic dictionary automatically, where this dictionary is served as gloss source for terms in the generated Amharic Sentiment Lexicon.

(e) Facebook Comments Data Set: This dataset consists of 2500 sentence/phrase level sentiment annotated facebook news users' comments collected from the Government Office Affairs Communication (GOAC) between 2008 and 2010. News that received high view counters/frequent comments were selected as "hot topics" and the associated comments labeled by professionals into either positive or negative sentiment.

(f) Amharic Web Corpus: This is an automatically crawled collection of web documents consisting mainly of news, politics and religion documents. It was collected and tokenized, automatically part-of-speech tagged and published under the HaBiT project [5]. The file is a single xml file of size 421 MB. It contains 20 million tokens. We will use this corpus for evaluation of the sentiment annotations.

3.2 Automatic Sentiment Score Calculation

To build an Amharic Sentiment Lexicon relying on bilingual dictionaries, we tried to transfer subjectivity or polarity information from polarity lexicon of resource rich language (i.e. English sentiment lexicon) to resource limited language (i.e. Amharic). Specific to this approach, an algorithm is developed to transfer labels from English SentiWordNet 3.0 and SO-CAL to Amharic Sentiment Lexicon.

As shown in Fig. 1 below, the Amharic-English bilingual dictionary is employed as a bridge between the two languages. The proposed approach of building Amharic Sentiment Lexicon is similar to Sinhala language in [2]. However, our approach is different at

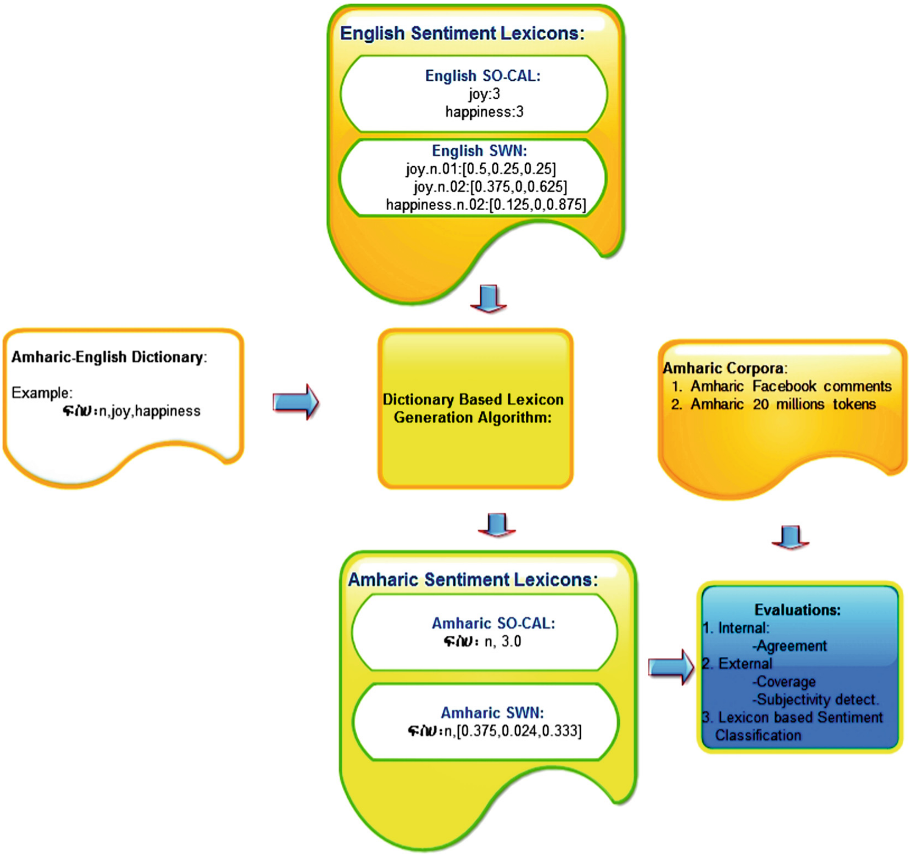
English Sentiment Lexicons:
English SO-CAL:
joy:3
happiness:3
English SWN:
joy.n.01:[0.5,0.25,0.25]
joy.n.02:[0.375,0,0.625]
happiness.n.02:[0.125,0,0.875]
Amharic-English Dictionary:
Example:
ፍስሀ፡n,joy,happiness
Dictionary Based Lexicon Generation Algorithm:
Amharic Corpora:
1. Amharic Facebook comments
2. Amharic 20 millions tokens
Amharic Sentiment Lexicons:
Amharic SO-CAL:
ፍስሀ፡ n, 3.0
Amharic SWN:
ፍስሀ፡n,[0.375,0.024,0.333]
Evaluations:
1. Internal:
-Agreement
2. External
-Coverage
-Subjectivity detect.
3. Lexicon based Sentiment Classification

Fig. 1. Dictionary based Amharic Sentiment Lexicon Building Framework

least in three features: (1) algorithm searches for each Amharic word as key rather than English words in the Amharic-English bilingual dictionary. That is why Amharic-English dictionary is used instead of English-Amharic dictionary. (2) our algorithm uses average weighting of synonymous of English to tokens to assign the sentiment score for the corresponding Amharic term in the bilingual Amharic-English dictionary and (3) the approach in this work will also handle negation words that appear in the Amharic text.

Consider Amharic word w_{a_i} in the Amharic-English Dictionary d_{ae} has corresponding meanings in English word(s), $w_{e\_i\_j}$. The Amharic Sentiment Lexicon is denoted by s_a, the English language sentiment lexicon by $s_{e.}$ As part of preprocessing, we normalized all Amharic words in the Amharic-English dictionary by replacing varied alphabets of the same sound with identical symbols. Moreover, a stemmer is applied during evaluations. Then the algorithm is represented:

```
BuildAmharicLexicon($s_e$)
Input: $s_e$ (EnglishSentiment Lexicon i.e. either SWN or SO-CAL)
Output: $s_a$ (generated Amharic Sentiment Lexicon)
            Format for $s_a$ is {$w_{ai}$ as key: scores, part of speech, gloss as list} pair.
Initialize: $d_{ae}$–load Amharic-English Dictionary from File and initialize other vars
  1.  For each $w_{a\_i}$ in the $d_{ae}$:
        a.  Search its {$w_{e\_i\_1}$, $w_{e\_i\_2}$,…, $w_{e\_i\_j}$,…, $w_{e\_i\_n-1}$, $w_{e\_i\_n}$}
        b.  For each synonym word $w_{e\_i\_j}$, search it in $s_e$
                  i.  If found, return all the polarity values(+,-,objective) of the
                      synsets
                 ii.  Keep them in a list L
        c.  Find length of +,-, objective list in L
        d.  Compute the average weighted polarity in L using equation (1)
        e.  Assign the average weighted polarity, part of speech, and gloss to $w_{a\_i}$
            in $s_e$. That is,
            $s_a[w_{a\_i}]$=[[Average weighted polarity], part of speech, gloss]

return $s_a$
```

Listing-1. Dictionary based Algorithm

Algorithm Description: In Listing-1, for each Amharic word in the bilingual dictionary, the aggregated sentiment values of corresponding English term(s) in SentiWordNet 3.0 (or SO-CAL) are assigned to the Amharic word if terms are found in the source lexicon. Finally, the Amharic Sentiment Lexicon contains each Amharic term, its average weighted polarity values, its corresponding part of speech and gloss definitions are assigned [8].

For each Amharic word w_{a_i} in the Amharic-English dictionary, there are list of English words($w_{e\_i\_1}$, $w_{e\_i\_2}$,. $w_{e\_i\_j}$., $w_{e\_i\_n-1}$, $w_{e\_i\_n}$) in the dictionary. The assumption is that the first English word $w_{e\_i\_1}$ has more closer in meaning to Amharic word w_{a_i} than the second word $w_{e\_i\_2}$ and in turn $w_{e\_i\_2}$ is more prominent to w_{a_i} than $w_{e\_i\_3}$ and so on. So more weight is given to the first word $\mathrm{w}_{\mathrm{e\_i\_1}}$ and least weight is given to the last word $\mathrm{w}_{\mathrm{e\_i\_n}}$. On the basis of this assumption, the weighted average sentiment of English words is computed and then assigned the resulting sentiment value to the corresponding Amharic word $\mathrm{w}_{\mathrm{a\_i}}$. Then, the sentiment score $(\mathrm{w}_{\mathrm{a\_i}})$ is given by:

$$\begin{aligned}
&= (\mathrm{score}(\mathrm{w}_{\mathrm{e\_i\_1}})\mathrm{X}(\mathrm{n}-1+1)+\mathrm{score}(\mathrm{w}_{\mathrm{e\_i\_2}})\mathrm{X}(\mathrm{n}-2+1)+..\,\mathrm{score}(\mathrm{w}_{\mathrm{e\_i\_j}})\mathrm{X}(\mathrm{n}-\mathrm{j}+1)..\\
&\quad+\mathrm{score}(\mathrm{w}_{\mathrm{e\_i\_n}})\mathrm{X}(\mathrm{n}-\mathrm{n}+1))/\mathrm{n}\\
&= \frac{\sum_{j=1}^{n} score_{eij}(w_{eij})x(n-j+1)}{n}
\end{aligned} \tag{1}$$

where n is the number of English words $\mathrm{w}_{\mathrm{e\_i\_1}}, \mathrm{w}_{\mathrm{e\_i\_2}},.., \mathrm{w}_{\mathrm{e\_i\_n-1}}, \mathrm{w}_{\mathrm{e\_i\_n}}$ and $\mathrm{score}_{\mathrm{eij}}$ is in turn refers to:

$$score_{eij} = \left\{ \begin{array}{ll} sentimentScore(w_{eij}), & if\ the\ lexicon \quad is \quad SO-CAL \\ sumOfSentimentScore(Synset(w_{eij}), & if\ the\ lexicon \quad is \quad SWN \end{array} \right\}$$

It should be noted that the gloss definitions are not found in Amharic English dictionary; rather it is propagated from another Amharic–Amharic dictionary. However, there are some terms without gloss definitions as they were not found in the Amharic–Amharic Dictionary [4].

The Amharic sentiment lexicons are generated and used under a few assumptions. (1) The senses of words in the bilingual dictionaries of the two languages are considered to be same. Thus, the sentiment score of an English word in sentiment analysis of English text is assumed to be identical to the sentiment score of an Amharic word in Amharic texts. (2) The parts of speech in both languages in the bilingual dictionary are assumed to be equivalent [2].

From English SentiWordNet 3.0 and SOCAL, two corresponding Amharic Sentiment lexicons are automatically generated. The sizes of these newly built Amharic lexicons are 13679 term entries and 5683 terms, respectively. The experimental results and evaluations are discussed in the subsequent section.

4 Results and Discussion

In this section, the experimental results relying on the generated Amharic sentiment lexicons are discussed. Each of the generated Amharic lexicons is evaluated in two ways using external data as well as intrinsic evaluation and mutual comparison of agreement. Finally the usefulness of the Amharic lexicons is tested in lexicon based sentiment classification of Amharic News Comments. To increase the string matching accuracy in generating the evaluation scores, we apply basic preprocessing: tokenization, stopword removal, punctuation mark removal and stemming.

(i) Tokenization: Amharic words are separated by space as delimiter. Thus, Amharic text is tokenized into words if space is found in between tokens.

(ii) Punctuation mark removal: Amharic punctuation marks such as ።(full stop), ፤ (double colon), ፣ (single quote), ፡ -(preface colon), ‹‹ ››(double quotes) and so on should be removed.

(iii) Normalization: Ethiopic script contains redundant symbols representing the same sound. These symbols are substituted by the symbol on the right. E.g. ('ሀ,ሃ,ሐ,ሓ,ኀ,ኃ,ኻ=>ሀ)፤ (ሰ,ሠ=>ሰ)፡ (ፀ,ጸ=>ፀ)፡ (ዐ,አ,ኣ,ዐ,ዓ=>ዐ)፡ (ቆ,ቈ,ቖ=>ቆ)፡ (ቁ,ቍ=>ቁ)፡(ኮ,ኰ=>ኮ)፡ (ጎ,ጐ=>ጎ)፡(ኋ,ዃ,ሗ=>ኋ)). Where, the arrow(=>) means "replaced by". That is, if either of the symbols on the left side found in the text, then it is replaced by the symbol right of the arrow(=>).

(iv) Stopword removal: For stopword removal, stopwords identified in [5] are used.

(v) Stemming: As Amharic is highly morphologically rich language, we develop a shallow stemmer by specifying some conditions to enforce and normalize the morphological variations of Amharic words to a common base form/stem. The stemmer is composed of a set of rules of the regular expressions to remove prefixes and suffixes. During stemming, if Amharic word has a match to one of the patterns of regular expressions, then the word is reduced to its corresponding stem by removing the matched suffix or/and prefix. For example, for the Amharic words such as 'ቆንጆው'/the one who is handsome/masculine/, 'ቆንጆውን' /someone who is handsome/masculine/, object form/, 'ቆንጆዎቹን'/those who are beautiful, ' ቆንጆዎች'/ those who are beautiful, 'ለቆንጆዋ'/to the beautiful one/feminine/and soon. Finally, these Amharic words are reduced to the common stem/root by converting the symbols to corresponding consonant of Amharic/'ቅ-ን-ጅ'/by our stemmer. More sample of Amharic stemmed words are attached in Appendix A of this report.

4.1 External to Lexicon Evaluation

In this research, external evaluation is carried out in two ways.

(a) Subjectivity detection: Each term in the lexicons is counted only if it is present in a particular Amharic news comments. If more terms of the lexicon found in the Amharic comment, then the comment is subjective, otherwise objective. The usefulness of automatically generated lexicons is evaluated by their accuracies in detecting subjectivity in Amharic texts. These lexicons are used to identify subjective (positive or negative) Amharic text from objective (neutral) Amharic text. The accuracy of subjectivity detection increases as we apply stemming along with the sentiment lexicons. As Amharic language is morphologically rich, the variations of words derivations are normalized to the same stem. Usually, the resulting stem will be in root form. This decreases the string mismatches which arise due to morphological variation of Amharic texts.

Table 1. Lexicons' subjectivity detection rate in percent on Facebook comments data

Sentiment Lexicons	Detection with No Stem (%)	Detection With Stem (%)
Amharic Manual (baseline)	43.23	93.56
Amharic SOCAL	31.28	96.65
Amharic SWN	75.83	99.33

Table 1 shows that the generated Amharic Sentiment SentiWord Net Lexicon outperforms the other lexicons, resulting in 99.33% correct subjectivity detection. The rate of detecting the subjectivity of Amharic news comments using Amharic Sentiment SO-CAL is, in turn, outperforming the manual Amharic sentiment lexicon (93.56%). This subjectivity detection result shows that each of the Amharic sentiment lexicons are outperforming above the subjectivity detection result of the baseline lexicon (the manually generated Amharic sentiment lexicon) by 5.77%. One of the reasons for this might be the size of SWN larger than the manual one. Thus, the degree to which the entries of SWN to appear in the Amharic news comment is higher than the other lexicons.

(b) Evaluation of lexicons by its coverage count (or in percent): This way of evaluating the generated lexicons is to measure the size of lexicons in terms of their coverage count in general corpus containing 20 millions tokens and 2500 Amharic facebook news comments.

Table 2. Lexicons' coverage (positive/negative count and in percent) on 2500 Amharic Facebook comments and 20 million tokens of Amharic web corpora

	2500 Amharic comments		20 millions Amharic tokens	
Lexicons	coverage(+,−) count	%	coverage(+,−) count	%
Manual	[4399, 2995]	31.95	[2713167, 2161501]	25.01
SOCAL	[5738, 3953]	41.87	[4169817, 3391213]	38.88
SWN	[9447, 4803]	61.57	[6645592, 4006072]	54.77

Discussion: The results in Table 2 verify that the coverage of the sentiment lexicon based on SWN is larger than the other lexicons on both data corpora. The other aspect that we can verify from this analysis is that the number of positive and negative opinion words are balanced in both SWN and SOCAL. However, the coverage of the Amharic lexicons in both corpuses is below the average benchmark semantic lexical coverage in other languages Welsh (25%) to Arabic (88%) [9], even though, it is very difficult to compare the coverage of Amharic sentiment lexicons with the coverage of these general purpose semantic lexical resources. Moreover, there is difference in languages intrinsic characteristics.

4.2 Internal to Lexicon Evaluation

Internal evaluation of the lexicon is the process of counting common positive and negative opinion terms in the two lexicons. The number of common terms in the two lexicons is expressed in percentage to show the extent of the agreement (overlap) between the generated lexicons and the manual lexicon (baseline). To evaluate the results of the proposed approach, we computed the agreement between lexicons.

Table 3. The Agreement (in percent) between Lexicons

	Amharic Sentiment Lexicons	Agreement (%)
1	SOCAL and Manual Lexicon	70.80
2	SWN and Manual Lexicon	59.49
3	SOCAL and SWN	66.40

Discussion: Table 3 depicts that the Amharic sentiment lexicon from SOCAL generated through dictionary approach has agreement or overlap of 70.80% with manual Amharic sentiment and it overlaps 66.40% with Amharic sentiment generated from SentiWord Net. On the other hand, Amharic Sentiment from SWN agrees for 59.49% of all terms with the manual Amharic Sentiment lexicon. Although the size of Amharic Sentiment from SWN is much larger (more than double) than Amharic Sentiment from SOCAL, the latter is more consistent in its sentiment scores with the other lexicons. The main purpose of finding agreement between lexicons is to know the extent to which the lexicons overlap.

The disagreement level of English Sentiment lexicons are compared in [10]. In this comparison, the agreement level of SWN is better with Harvard General Inquirer (77%) than other English Sentiment lexicons (MPQA, Opinion Lexicon, LIWC). Amharic generated sentiment lexicon is below the agreement levels reported for the English language resources. The reason for this is that it is very difficult to compare Amharic Lexicon with English lexicon as the two languages are very different in morphology apart from difference in cultural connotations. So, as Amharic is morphologically rich where there could be more variations of terms in the lexicons than the terms in English lexicons. This might decrease the extent of agreement of the Amharic sentiment lexicons.

4.3 Lexicon Based Sentiment Classification

Besides the evaluations in the earlier subsections, we will also evaluate the usefulness of the generated lexicons and their combinations for sentiment classification of Amharic facebook news comments. Prior to sentiment aggregations of Amharic texts, we apply basic text preprocessing (tokenization, punctuation mark removal, normalizing Amharic script symbols, stopword removal, spelling corrector, stemming, etc.). The effect of stemming and negation detection technique on Amharic text is investigated to increase the accuracy of lexicon based Amharic sentiment classification.

Table 4. The accuracy (in percent) of Lexicons for Sentiment Classification

Amharic Senti.Lexicons	Accuracies (%)		
	NoStem+NoNeg.	*Stem+NoNeg.*	*Stem+Neg.*
Manual (baseline)	16.7	42.9	52.7
SOCAL	14.6	46.3	50.8
SWN	30.9	50.1	54.7
Manual + SOCAL	37.2	63.4	72
Manual + SWN	49.9	65.9	74.6
SOCAL + SWN	43.7	66.6	73.5
Manual +SOCAL + SWN	53.7	75.8	**86.2**

Discussions: The usefulness of lexicons is evaluated in terms their accuracies of classifying sentiment of Amharic facebook news comments as shown in Table 4. In general, the results in Table 4 also reveal the effect of applying stemming and negation handling on Amharic texts to boost the performance of sentiment classification. The automatically generated lexicon from English SWN outperforms the performance of the manual (baseline) for classifying sentiment of Amharic texts. On the other hand, the manual lexicon (with stemming and negation handling) in turn outperforms the automatically generated lexicon from English SO-CAL. However, Amharic sentiment lexicon (with stemming and without negation handling) from English SO-CAL relatively performs better than the Amharic manual sentiment lexicon in classifying sentiment of Amharic texts. The combination of the automatically generated sentiment lexicon (SWN) with Manual sentiment lexicon (baseline) outperforms (with accuracy of 74.6%) the other combinations of lexicons for sentiment classification of Amharic texts. The automatically generated lexicons (SO-CAL + SWN) perform well (with accuracy of 73.5%) for classification sentiment in Amharic texts. Yet another combinations of the three lexicons(SO-CAL + SWN + Manual) outperform (with accuracy of 86.2%) the other lexicons or their combinations.

4.4 Error Analysis

Unfortunately, it is challenging to trace the sources of errors in the automatically generated lexicons. Let us try to point out some of the causes for the subsets of generated errors in the automatically translated Amharic sentiment lexicons. We present manually identified errors in this subsection:

(a) On-spot analysis of errors detected in lexicons: In the automatically generated lexicons, the generated errors in sentiment transfer from source language terms to target language terms are mainly caused by the following issues: (i) We identified mistranslation of Amharic terms in SWN by the bilingual dictionary. For example, in Amharic SWN, the word ሸጋ(“SHEGA”) is incorrectly translated into target ‘hibernation’ and it is wrongly assigned with negative sentiment. But, the correct meaning of the word ሸጋ(“SHEGA”) means “nice” which has positive sentiment value. (ii) We got correctly translated terms which are assigned opposite in polarity to source terms. For example, in Amharic SWN, the word ቆራጥ(“QORAT”) correctly translated into’courageous decisive in manner’. The translation is correct, but wrong sentiment value (opposite) is assigned. (iii) We discovered few terms are correctly translated, the same sentiment polarity but different sentiment strength. For example, ሰይጣን(“SEYTAN”) means “devil” in Amharic SWN. It is correctly translated and assigned negative sentiment value but the sentiment strength is small. (iv) We found some terms in SWN which are correctly translated but different sentiment due to cultural connotations. For example, the term እርካሽ(“ERKASH”) means “cheap” and it is assigned positive sentiment. In English, this might be correct, but in Amharic the word እርካሽ(“ERKASH”) has negative connotation in that it refers to something which is sinfulness and lower in quality. The level of these type errors can be minimized by incorporating context dependent lexicon generated from Amharic corpus. This leads to another venue of future researches.

(b) Analysis of Incorrectly Detected Amharic News Comments: We identified the associated reasons why the sentiment/subjectivity of Amharic News Comments are wrongly detected. The subsets of reasons which cause wrong subjectivity/sentiment detection include: (i) We discovered some Amharic comments which contains sarcasm and idioms. The nature of these comments is difficult to detect its sentiment relying on ordinary lexicon. For example, 'የራሷ አሮባት የሰው ታማስላለች/solve your own problem before talking about others/is an Amharic proverb that cannot be translated directly relying on ordinary dictionary. That is why it is wrongly classified by the sentiment lexicons. (ii) Besides handling negation, we discovered that further formulations of linguistic rules (e.g. intensifiers, contrast rules, conjunction rules) are required to handle wrongly detected Amharic news comments. For example, 'ኃይሌ ገ/ስላሴ በሩጫ ጎበዝ እንጂ ስለኢትዮጵያዊ መንፈስ ሊሰበክ አይችልም'/H/G/Silasie is the best runner, however, he cannot preach about Ethiopianism/. In this example, the sentiment computation is failed as the semantic orientation of the text is diverted by contrasting word እንጂ /however/. The text next to this contrasting word has dominant sentiment than the phrase before it. (iii) We identified some Amharic comments which are wrongly annotated by human annotator. For example, 'በጣም አሳዛኝ ነው'/it is very tragedy/. The sentiment of this comment is labeled wrongly as positive by the human annotator. Thus, the labeled data should be reviewed for correction. (iv) We also found some Amharic news comments which are detected wrongly as their meanings are implicit where their interpretations are connected to pragmatics. Such context dependent text connotations are difficult to handle by explicitly using our generated lexicons. For example, the Amharic news comments ‘‘ፍትህ ለአማራ'’/Justice to Amhara/is assigned negative sentiment by annotator, but lexicon based classifier wrongly detected it as positive. The reason is that the lexicon based classifier is limited in handling the contextual meaning of the comment in place. This comment is primarily connected to the context and implicitly associated to the meaning behind the original news post.

5 Conclusions and Recommendations

Amharic sentiment lexicon is one of the resources required for Amharic sentiment analysis. Yet, extensive lexical resources are expensive to build. To remedy this problem, we propose a dictionary based approach for generating Amharic Sentiment lexicon. It requires a bilingual dictionary to propagate polarity information from sentiment lexicon of source language to target language. This approach can be used to generate large scale sentiment lexicons. However, it generates general sentiment lexicons that may lack accuracy for sentiment analysis. It is unable to handle cultural and language specific connotations in a particular language. So to address these issues, we proposed another approach which is a corpus based approach to be done in the forthcoming work in our project. Then once processed and refined, dictionary based approach generates two Amharic lexicons: Amharic Sentiment Lexicon_SOCAL(5683) and Amharic Sentiment Lexicon SWN(13679).

The lexicons generated using dictionary based approach are evaluated by agreement (internal), coverage (external), subjectivity detection rate (external) and its performance in lexicon based Amharic text sentiment classification. The Amharic sentiment lexicon generated from SWN is not only good in internal evaluation (i.e. it has acceptable agreement rate with both the manual and SOCAL lexicons) but it also has higher coverage in both test corpora than the other two sentiment lexicons. Moreover, the lexicon based sentiment classification of Amharic Sentiment lexicon from SWN outperforms the other Amharic Sentiment lexicon(including the baseline).

This work demonstrates that it is possible to automatically generate sentiment lexicons relying on available bilingual dictionaries to minimize time and labor cost of manual sentiment lexicon preparation. The generated lexicon expected to get sufficient sentiment lexicon size by porting from resource rich language. Some of the contributions of this work are briefly summarized below:

- Being an automatic approach, the algorithm developed reduces cost and time of labeling terms in sentiment lexicon.
- The approach developed is generic enough that it can be adapted to generate sentiment lexicon to other resource limited languages.
- The entries in our lexicons contain part of speech information that can be applied or utilized in other linguistic tasks of interest.
- The approach can also be adapted to other tasks of natural language processing including information extraction, multilingual semantic lexicons, question and answering, just to name a few.
- The combination of all generated lexicons with the manual lexicon achieves best sentiment classification performance on Amharic texts.
- The code and related resources will be available online for research communities.

Yet, the generated sentiment lexicons may lack accuracy for sentiment analysis of a particular language context where the approach potentially does not sufficiently capture the cultural and language specific connotations in a particular language. To address these issues, we may need to consider corpus based approaches that capture and incorporate semantic information on the specific meaning of a term in a given context to provide higher precision in sentiment score assignment for each particular instance.

Appendix A. Sample of Amharic Stemmed Words and Its Variant Word Forms

Three sample words are selected. These are from verb, noun and adjective categories. The base forms of these sample words include ሰበረ//seBeRe/means ‘he break something’, ቤት/Beet/means ‘home’ and ቆንጆ/qonJo/means ‘beautiful’, respectively. Table 5 below shows the different forms of these words and their corresponding stems or roots.

Table 5. Sample of the different variant word forms and their corresponding stems.

Surface Word	SERA form	Stem	Root
'ሰበርኩ'	seBeRku	'ስብር'	'ስብር'
'ሰበርክ'	seBeRk	'ስብር'	'ስብር'
'ሰበርሽ'	seBeRX	'ስብር'	'ስብር'
'ሰበረ'	seBeRe	'ስብር'	'ስብር'
'ሰበረች'	seBeRec	'ስብር'	'ስብር'
'ሰበርን'	seBeRn	'ስብር'	'ስብር'
'ሰበራችሁ'	seBeRachu	'ስብር'	'ስብር'
'ሰበሩ'	seBeRu	'ስብር'	'ስብር'
'ሰበርኩት'	seBeRkut	'ስብር'	'ስብር'
'ሰበርከው'	seBeRkeW	'ስብር'	'ስብር'
'ሰበርሽው'	seBeRXW	'ስብር'	'ስብር'
'ሰበረው'	seBeReW	'ስብር'	'ስብር'
'ሰበረችው'	seBeRecW	'ስብር'	'ስብር'
'ሰበርነው'	seBeRneW	'ስብር'	'ስብር'
'ሰበራችሁት'	seBeRachut	'ስብር'	'ስብር'
'ሰበሩት'	seBeRut	'ስብር'	'ስብር'
'እሰብራለሁ'	IseBRaLehu	'ስብር'	'ስብር'
'ትሰብራለህ'	tseBRaLeh	'ስብር'	'ስብር'
'ትሰብሪአለሽ'	tseBRiaLeX	"ትሰብሪ'አለ"	'ስብር'
'ይሰብራል'	YseBRaL	'ስብር'	'ስብር'
'ትሰብራለች'	tseBRaLec	'ስብር'	'ስብር'
'እንሰብራለን'	InseBRaLen	'ስብር'	'ስብር'
'ትሰብራላችሁ'	tseBRaLachu	'ስብር'	'ስብር'
'ይሰብራሉ'	YseBRaLu	'ስብር'	'ስብር'
'አልሰብርም'	aLseBRM	'ስብር'	'ስብር'
'አልሰበርኩም'	aLseBeRkuM	'ስብር'	'ስብር'
'አትሰብርም'	atseBRM	'ስብር'	'ስብር'
'አትሰብሪም'	atseBRiM	'ስብር'	'ስብር'
'አንሰብርም'	anseBRM	'ስብር'	'ስብር'

'አትሰብሩም'	atseBRuM	'ስብር'	'ስብር'
'አይሰብርም'	aYseBRM	'ስብር'	'ስብር'
'አልሰበሩም'	aLseBeRuM	'ስብር'	'ስብር'
'አልሰበረችም'	aLseBeRecM	'ስብር'	'ስብር'
'አይሰብሩም'	aYseBRuM	'ስብር'	'ስብር'
'ሰባበሩ'	seBaBeRu	'ስብር'	'ስብር'
'ሰባበርን'	seBaBeRn	'ስብር'	'ስብር'
'ሰባበረች'	seBaBeRec	'ስብር'	'ስብር'
'ሰባበርሽ'	seBaBeRX	'ስብር'	'ስብር'
'ሰባበርህ'	seBaBeRh	'ስብር'	'ስብር'
'ሰባበሩ'	seBaBeRu	'ስብር'	'ስብር'
'ሰባበረ'	seBaBeRe	'ስብር'	'ስብር'
'ሰባበርኩ'	seBaBeRku	'ስብር'	'ስብር'
'አልሰባበሩም'	aLseBaBeRuM	'ስብር'	'ስብር'
'አልሰባበረችም'	aLseBaBeRecM	'ስብር'	'ስብር'
'አልሰባበርንም'	aLseBaBeRnM	'ስብር'	'ስብር'
'አልሰባበረም'	aLseBaBeReM	'ስብር'	'ስብር'
'አልሰባበርክም'	aLseBaBeRkM	'ስብር'	'ስብር'
'አልሰባበርንም'	aLseBaBeRnM	'ስብር'	'ስብር'
'አትሰባብሩትም'	atseBaBRutM	'ስብር'	'ስብር'
'አንሰባብረውም'	anseBaBReW	'ስብር'	'ስብብር'
'አይሰባብሩትም'	aYseBaBRutM	'ስብር'	'ስብር'
'ያልሰበረ'	YaLseBeRe	'ስብር'	'ስብር'
'ያለሰበሩ'	YaLeseBeRu	'አለሰበር'	'ስብር'
'ያለሰበረች'	YaLeseBeRec	'አለሰበረች'	'ስብር'
'ያልሰባበረ'	YaLseBaBeRe	'ስብር'	'ስብር'
'ሳትሰባብረ'	satseBaBRe	'ስብር'	'ስብር'
'ሳንሰብረ'	sanseBRe	'ስብር'	'ስብር'
'ሳንሰባብር'	sanseBaBR	'ስብር'	'ስብር'
'ሳልሰባብር'	saLseBaBR	'ስብር'	'ስብር'

'ሳልሰብር'	saLseBR	'ስብር'	'ስብር'
'ሳይሰበር'	saYseBeR	'ስብር'	'ስብር'
'እየሰባበሩ'	IYeseBaBeRu	'ስብር'	'ስብር'
'ቤቴ'	Beetee	'ቤት'	'ብት'
'ቤትህ'	Beeth	'ቤት'	'ብት'
'ቤታችሁ'	Beetachu	'ቤት'	'ብት'
'ቤታችን'	Beetacn	'ቤት'	'ብት'
'ቤትሽ'	BeetX	'ቤት'	'ብት'
'ቤትሽን'	BeetXn	'ቤት'	'ብት'
'ቤቱ'	Beetu	'ቤት'	'ብት'
'ቤቱን'	Beetun	'ቤት'	'ብት'
'ቤታቸው'	BeetaceW	'ቤት'	'ብት'
'ቤታቸውን'	BeetaceWn	'ቤት'	'ብት'
'ቤታችሁን'	Beetachun	'ቤት'	'ብት'
'ቤታችሁ'	Beetachu	'ቤት'	'ብት'
'ቤትዎ'	BeetWo	'ቤት'	'ብት'
'ቤትዎን'	BeetWon	'ቤት'	'ብት'
'ቤቶች'	Beetoc	'ቤት'	'ብት'
'ቤቱ'	Beetu	'ቤት'	'ብት'
'ቤቱን'	Beetun	'ቤት'	'ብት'
'የቤቱ'	YeBeetu	'ቤት'	'ብት'
'ለቤቱ'	LeBeetu	'ቤት'	'ብት'
'ቤት'	Beet	'ቤት'	'ብት'
'ቤትን'	Beetn	'ቤት'	'ብት'
'የቤት'	YeBeet	'ቤት'	'ብት'
'ለቤት'	LeBeet	'ቤት'	'ብት'
'ቤቶቹ'	Beetocu	'ቤት'	'ብት'
'ቤቶቹን'	Beetocun	'ቤት'	'ብት'
'የቤቶቹ'	YeBeetocu	'ቤት'	'ብት'
'ለቤቶቹ'	LeBeetocu	'ቤት'	'ብት'

'ቤቶች'	Beetoc	'ቤት'	'ብት'
'ቤቶችን'	Beetocn	'ቤት'	'ብት'
'የቤቶች'	YeBeetoc	'ቤት'	'ብት'
'ለቤቶች'	LeBeetoc	'ቤት'	'ብት'
'ቆንጆው'	qonJoW	'ቆንጆ'	'ቅንጅ'
'ቆንጆውን'	qonJoWn	'ቆንጆ'	'ቅንጅ'
'ቆንጆዎቹን'	qonJoWocun	'ቆንጆ'	'ቅንጅ'
'ቆንጆዎች'	qonJoWoc	'ቆንጆ'	'ቅንጅ'
'ለቆንጆዋ'	LeqonJoWa	'ቆንጆ'	'ቅንጅ'

References

1. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
2. Medagoda, N., Shanmuganathan, S., Whalley, J.: Sentiment lexicon construction using sentiwordnet 3.0. In: Proceedings of the 11th International Conference on Natural Computation (ICNC), pp. 802–807, IEEE (2015)
3. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Language Resources and Evaluation (LREC), vol. 10, pp. 2200–2204 (2010)
4. Lexical Data Repository of the Ge'ez Frontier Foundation. https://github.com/geezorg/data. Accessed 15 Feb 2017
5. Project teams from Addis Ababa University, Masarykova univerzita, Norges teknisk-naturvitenskapelige universitet, The University of Oslo, Hawassa University, 7F14047 HaBiT - Harvesting big text data for under-resourced languages. http://habit-project.eu/wiki/HabitSystemFinal. Accessed 05 May 2018
6. Gebremeskel, S.: Sentiment mining model for opinionated Amharic texts, Unpublished Masters thesis, Department of Computer Science, Addis Ababa University, Addis Ababa (2010)
7. Tilahun, T.: Linguistic localization of opinion mining from Amharic blogs. Int. J. Inf. Technol. Comput. Sci. Perspect. **3**(1), 890 (2014)
8. Denecke, K.: Using SentiWordNet for multilingual sentiment analysis. In: Data Engineering Workshop. pp. 507–512, IEEE (2008)
9. Piao, S., Rayson, P.: Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In: European Language Resources Association (ELRA), pp. 2614–2619 (2016)
10. Christopher Potts: Sentiment Symposium Tutorial Lexicons, Stanford Linguistics. http://sentiment.christopherpotts.net/lexicons.html. Accessed 10 Jan (2019)