A Machine Learning Approach to Multi-Scale Sentiment Analysis of Amharic Online Posts

Wondwossen Philemon
HiLCoE, Computer Science Programme, Ethiopia
wondaphil@yahoo.com

Wondwossen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondisho@yahoo.com

Abstract

Sentiment analysis classifies opinions as positive and negative based on polarity of words. Recent development in the field mainly focuses on online posts on social media, product marketing and news sites. In this paper, we present a multi-scale sentiment analysis model for Amharic using supervised machine learning approach.

Despite the growing availability of electronic Amharic text in the web, there are no sufficiently available sentiment corpora to be used for research. We developed our own corpus by collecting around six hundred posts from online sources. Even the scarce texts available on the web are mostly transliterated in Latin due to lack of accessibility and convenience to typing. Besides, Amharic belongs to the Semitic family of languages which are characterized by morphological complexity which resulted in data sparseness. Thus, preprocessing techniques have been employed to clean the data, to convert transliterations to the native Ethiopic script and to change words to their base form by removing the inflectional morphemes. After preprocessing, the corpus is manually annotated by giving polarity and sentiment intensity scale values. This multi-scale approach to sentiment analysis provides a more refined breakdown than the traditional positive-negative binary scheme. The approach is preferred in cases where comparison and ranking of opinions is vital. We employed Naïve Bayes machine learning algorithm and used unigram, bigram and hybrid variants as features.

The experiment results show that, among the three learning setups, the accuracy of the bigram approach is found to be promising. Particularly, for the intensified positive and negative polarity classes, the bigram approach performed better. Generally, the results are encouraging despite the morphological challenge in Amharic, the data cleanness and small size of data. We are convinced that the results could improve further with a larger corpus.

Keywords: Ethiopic; Multi-scale; Sentiment Analysis; Unigram; Bigram; 4343 Naïve Bayes; Machine Learning

1. Introduction

In this information and technology age, a huge amount of textual information is readily available which require systematic and efficient means of categorization. Although much of the categorization works had been focusing on the basis of subject matter, the rapid growth of social media and blogs which inherently express personal positive or negative opinions about a given topic demanded a different kind categorization [1]. This new field of categorization, called Sentiment Analysis, is the concern of this paper.

People often use online services to share their feelings and comment on other people's posts, newly released products or current issues. Analysing sentiments is critically important for governments, manufacturers, and companies because it allows them to get a feedback and general review of clients' and customers' feelings on their products and services [2]. Customers also require these services as they are influenced by online reviews and

recommendations while purchasing products and services [3, 4].

A multi-scale sentiment analysis which indicates the extent to which a given text is positive or negative gives a better insight to sentiment analysis. Interpretation of opinions could be challenging for humans as binary distinction of opinions as only positive or negative may not suffice [1, 2, 5]. The scale values show the strength of positivity or negativity of a text as rank. It provides a quick indication of the tone of a text and provides a more refined analysis which is important for several real life applications such as comparison of several opinions and for giving ranks to different opinions.

Although sentiment analysis can be applied to any human language, some of the approaches could be language specific. The majority of sentiment analysis studies that have been conducted are for English language and the methods cannot be directly implemented on other languages. Languages in the Semitic family that are typically rich in regard to morphology are not exceptions in this regard [3]. The root-template morphology that characterizes Semitic languages results in high degree of word productivity and hence data sparseness which demanded natural language processing (NLP) preprocessing techniques to analyze sentiments [6, 7].

Amharic which is also from the Semitics branch is the official working language of Ethiopia spoken by about 30 million people. However, it is one of the least researched and under resourced languages with relatively few computational linguistic works. Users have been using Latin alphabet to transliterate Amharic texts and express their view in the social media. Recently, the introduction of Amharic Unicode font and its integration in different platforms (including smart devices) has paved a way for many online publishers and users to interact using the native Ethiopic script. This trend of Amharic being used as a medium of online communication among speakers is contributing to the enrichment of the web with Amharic contents thereby opening further study opportunities for the language.

In this paper, an attempt has been made to apply supervised machine learning approach for sentiment analysis on Amharic online posts which are written with Ethiopic script. The main focus is on comments, reviews and feedback posts which are typically brief and reflect opinions and personal feelings. We addressed the challenges of applying machine learning approach to the task and the feasibility of multi-scaling. We have also explored the impact of morphological analysis to sentiment analysis on Amharic by integrating the system with a morphological analyzer tool.

2. Related Work

There are a number of studies on sentiment analysis and a few are closely related to our study. We focused on two aspects for selecting studies for our discussion, namely the language family (Amharic and other morphologically rich languages such as Arabic) and the approaches followed (term-counting versus machine learning).

2.1 Sentiment Analysis on Semitic Language

From the Semitic family, Arabic has a number of NLP studies and attempts. The different computational linguistic resources for Arabic like corpora and tools for tokenization, Part-of-speech (POS) tagging, morphological analysis, stemming and machine translation put the language up front in the family. Besides, the web has relatively richer Arabic content than other Semitic languages [6, 7].

Two papers on Arabic by Abdul-Mageed *et al.* [6] and Mourad and Darwish [7] discuss in detail the peculiar aspects of morphological complexity. The studies employed morphological analysis on their datasets. Both studies were on subjectivity and sentiment analysis (SSA) that initially classifies documents as subjective and objective which further proceeds to polarity classification.

We encountered only one sentiment analysis study on Amharic language by Gebremeskel [8]. The experimental results, as reported by the author, indicate that the rule based model they developed performs well. However, it can also been noted that

the performance would have been improved if morphological analysis has been employed.

2.2 Term Counting Approach

Gebremeskel [8] followed the term counting approach to sentiment analysis. In the model he proposed, once sentiment words and valence shifter words are detected using a sentiment lexicon, terms will be assigned weight by considering the effect of diminisher, intensifier or negation terms. The total polarity weight of a review is calculated by adding the polarity weight of individual sentiment terms.

Although the term counting approach may be considered as a valuable alternative for underdeveloped languages like Amharic which are facing challenges in building a corpus, systems developed using this approach are not easy to scale-up. Besides, machine learning performs with less human intervention.

2.3 Machine Learning Approach

Abdul-Mageed *et al*. [6] followed a supervised machine learning approach which employs Support Vector Machine (SVM) algorithm for classification. The sentences they collected were manually annotated by two native speakers and the Kappa (K) inter-annotator agreement was also computed to evaluate the judgment of the annotators. They employed word form and POS tagging as features. To minimize data sparseness, they converted all words to the corresponding base forms. Their experimental results showed that morphological analysis has more influence on sentiment than subjectivity analysis. However, use of POS tags as features practically had no effect in enhancing the accuracy. Their subjectivity analysis study has pointed out a research gap for different languages such as Amharic. Their study has also made it apparent that morphological analysis has an impact on sentiment analysis of Semitic languages.

Mourad and Darwish [7] also researched SSA on Arabic and followed machine learning approach. Their corpus is mainly composed of twits and the annotators worked together to resolve any disagreement rather than measuring inter-annotator agreement. Letter normalization is also performed to convert variant forms of letters to a single form. They faced challenges while dealing with smileys and emoticons which are commonly interchanged due to the right-to-left writing system of Arabic. The other challenge relates to dialects which normally lack spelling standards and introduce many new words into the language.

Mourad and Darwish used more number of features like POS, n-gram, tweets-specific features, presence of emoticons, usage of decorating characters, punctuations, elongations and repetitions. They implemented Naïve Bayes and SVM classifiers using NLTK tool but the former is found to perform better. They claimed that their experimental results surpassed all previous Arabic SSA studies. The machine translation technique used by Mourad and Darwish to translate twits written in English to Arabic is a very good solution for tackling the problem of lack of corpora for underdeveloped languages.

3. The Proposed Sentiment Analysis Model

The sentiment analysis model we proposed is depicted in Figure 1. It has to be noted that the model has a number of components for preprocessing, training a classifier and classifying an input post. There are two major lines of flows for preprocessing of the corpus and an input post whose outputs are annotated lemmatized training data and lemmatized post, respectively. The classifier which is trained on the annotated lemmatized training data classifies the lemmatized version of the post and gives its multi scale polarity value. Most of the components are commonly available in other sentiment analysis models but the Latin transliteration converter and the morphological analysis components are peculiar to our model. The upcoming sections describe the various components of the model.

3.1 Preparation of Corpus

We were able to collect 608 posts from Facebook, Twitter, DireTube and Ethiopian Reporter websites.

This is relatively small as compared to previous studies [1, 6, 9].

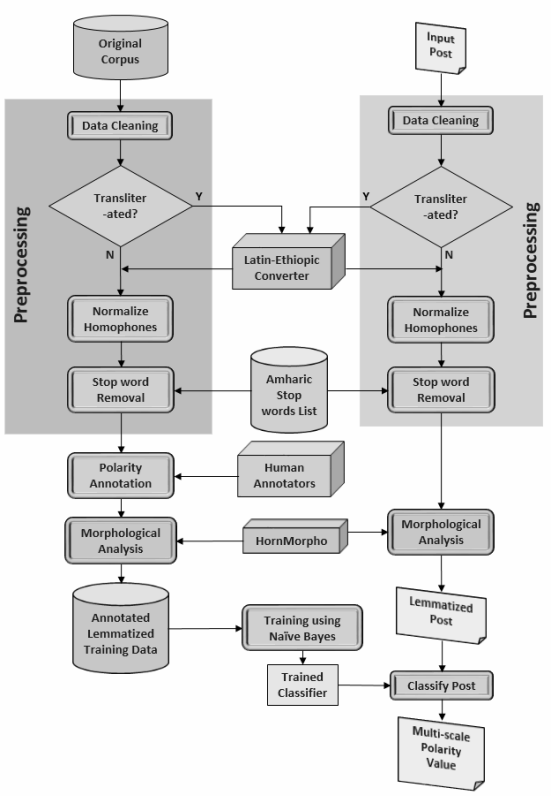

Figure 1: The Proposed Amharic Machine Learning Sentiment Analysis Model

However, data extraction for Amharic was difficult due to lack of Amharic web content and it took a significant proportion of time. Some of the posts are transliterated in SERA (System for Ethiopic Representation in ASCII) and later converted to the Ethiopic Unicode format. Correction of spelling and grammatical errors and removing irrelevant contents such as Tweeter hashtags were done manually.

Before the actual sentiment classification is done, a number of preprocessing tasks are performed. The first task is data cleaning. Then any Latin transliterated text is converted to Ethiopic script to maintain data uniformity and consistency. Normalization is another preprocessing which converts homophone variations of Amharic writing to a common symbol. For instance, the letters '**ሃ**', '**ሐ**', '**ሓ**', '**ኃ**', '**ኻ**' are converted to the form '**ሀ**' due to the fact that all have the same sound "ha".

Finally, we removed stop words from the data set as they are assumed to have less significance for analyzing sentiment while their frequency count dominates all other words. As far as we know, there is no Amharic stop words list available for use and we manually prepared the list using the training data and other Amharic texts as a source. All words in the prepared stop word are lemmatized and normalized for data uniformity.

3.2 Annotation of Corpus

In supervised machine learning, annotation of the training corpus with the target sentiment output is necessary. Each of the sentences in the collection is annotated according to its sentiment polarity as well as sentiment strength. We adopted a two scale scheme which scale positive sentiments further as +1 and +2 for less positive and more positive posts, respectively, as well as negative polarities as -1 and -2 for less negative and more negative posts, respectively. The neutral sentences were annotated as 0. We have limited our scale to these five ranks because of the high degree of productivity of Amharic words that makes it difficult to have a clear distinction between polarity strengths of texts. Besides, we strongly believe that the limited corpus size would limit the available scale options that we can practically employ.

The annotation was done independently by two native speakers of Amharic as suggested by previous researchers and the conformity between the annotators was statistically measured [5, 6, 7]. We found the value of the Kappa (K) parameter to be 0.823 (82.3%) which is an evidence that there is a very good strength of agreement between the two annotators.

3.3 Lemmatization

This preprocessing task involves morphological analysis which converts every word of the post to their base forms to avoid data sparseness. The task is done by integrating our system with HornMorpho - a freely available Python tool for morphological analysis and POS tagging of Amharic text [10]. HornMorpho analyses a given word in detail and we parse the output from HornMorpho using regular

expressions to pick the base form. A post is tokenized, each word is converted to its base form and finally the post is reconstructed with these new forms.

Some Amharic words have a negation affix which is stripped off during lemmatization. For instance, the word "አያምርም" (meaning "not attractive") has a base form of "አማረ" (meaning "becomes attractive") which no longer has a negative polarity. In such cases, the parser sets a Boolean variable - polarity reverser - which indicates that polarity value has to be reversed in later stages to maintain the polarity.

3.4 Model Learning

For the model we presented in Figure 1, the features we have used are n-grams (unigrams and bigrams). Since we have a limited sized corpus, we did not employ trigrams but we also used a hybrid of unigram and bigram to improve performance [1]. Even though it is suggested that using POS tag as a feature improves performance, we were not able to implement that because the morphological analyzer we have employed, HornMorpho, considers adjectives which are dominant sentiment carrying word classes as nouns [12]. Besides experiments by Abdul-Mageed *et al*. [6] indicated that POS tagging only favors subjectivity analysis, rather than sentiment analysis.

We train the Naïve Bayes classifiers on the entire corpus and the classifier accepts a given post to classify it according to the classification knowledge acquired on training. During the experiment, however, the corpus is divided into training set (80%) and test set (20%) and the classifier is trained with the training set only.

3.5 System Prototype

The system prototype has been developed using Python programming language where the implementation includes the preprocessing, lemmatization, training, classification and evaluation components discussed in the preceding sections. We used the popular Natural Language Toolkit (NLTK) which provides automatic classifiers like the Naïve Bayes and the tkinter module which allows development of a window-based interface [6].

When the system prototype is run, loading of the corpus and training are done only once during startup to make it more efficient while running. The main window accepts an Amharic text in the Ethiopic Unicode format or SERA transliteration. The user can then select the options to classify the input using either unigram or bigram approaches. A sample input text entered and the resulting classification given by the unigram classifier is presented in Figure 2.

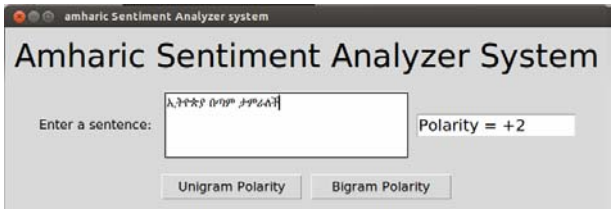

Figure 2: Sample Input Text and the Polarity Value

In addition to displaying the polarity value in the main window, the system prototype also displays detailed analysis result in Python shell which shows the input post, its lemmatized version, the value of the polarity reverser variable (discussed in Section 3.3) and the actual polarity value.

In addition to classifying a post, the system allows the user to evaluate the three language models, namely unigram, bigram and the hybrid model. The detailed evaluation result displayed in the Python shell constitutes the total number of posts in the corpus, proportion of posts used for training and testing, the list of available polarity classes, the accuracy, the three evaluation metrics for each polarity class (in tabular form), and the list of top features. A sample of such analysis result is presented in Figure 3.

The prototype we developed can run in different platforms but it is expected to correctly analyse only shorter length posts because the training corpus is composed of relatively shorter posts with three to four words on the average. Besides, Amharic dialects, slangs and abbreviations may not be handled fully as the corpus has little or no variations of such forms.

```
BIGRAM analysis:
Total number of sentences: 608
Number of training set: 456
Number of test set: 152
Labels/Tags: ['-2', '+1', '0', '-1', '+2']
Accuracy: 0.441
         Precision    Recall   F-measure
-2          0.5         0.29      0.367
-1          0.38        0.07      0.118
0           0.17        0.04      0.065
+1          0.42        0.9       0.573
+2          0.9         0.47      0.618
Top features: [('ነው', 'በረታ'), ('እጅግ', 'በጣም'),
('ጥሩ', 'ነው'), ('ደስ', 'አለ'), ('እጅግ', 'እጅግ'),
('ምኞት', 'ነው'), ('ሆነ', 'ቻለ'), ('አዲስ', 'አበባ'),
('ምሽት', 'ሆነ'), ('አለ', 'ነው')]
```

Ln: 6300 Col: 0

Figure 3: Sample Output of Bigram Evaluation

4. Result and Discussion

4.1 Evaluation Procedures

The experiment is done to measure the overall performance of the developed supervised machine learning sentiment analysis model. Out of the available corpus, the portion of the training set is 486 posts and the remaining 122 posts are set aside as test data set. Each of the posts in the test set is used as an input post one by one and the system returns their polarity. These set of observed polarity values are compared to the annotated polarity values which is our reference set. For a given distribution of the training and test sets, we run the experiment for unigram, bigram and the hybrid models and recorded the results. The results achieved are presented in the forthcoming section.

4.2 Evaluation Results

The most commonly used evaluation metrics in sentiment analysis are accuracy, precision, recall and F-score [3, 6]. There are five polarity classes in our case (‘-2’, ‘-1’, ‘0’, ‘+1’, and ‘+2’) and we calculated precision, recall and f-measure for each class but accuracy is calculated for the classifier as a whole. The evaluation result of each metrics for each language model in the corresponding classes is presented in Table 1.

4.3 Interpretation of Results

We have observed that the classifier’s accuracy is better for the bigram language model which suggests that considering a window of two words while determining multi-scale sentiment is recommended. Although experimental results by Pang *et al.* [1] suggested that the hybrid model performs even better than bigram, the accuracy is the lowest in our case. Precision and recall for unigram and hybrid models are also very close in all polarity classes.

In all language models, precision and recall are higher for the ‘+1’ polarity class because nearly 40% of the posts in our corpus belong to this class. Particularly in the bigram model, precision is significantly very high for the ‘-2’ and ‘+2’ classes. Obviously, this is due to the effect of the intensifiers and diminishers which are mostly accumulated here and the bigram captures these valence shifters more precisely. We repeatedly performed the experiment and these findings remained unchanged. Generally, the values of the evaluation metrics are encouraging and more training data is assumed to improve the performance.

Table 1: Evaluation Results of the Three Language Models

Polarity Class	*Unigram*			*Bigram*			*Unigram-Bigram*		
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F- score*	*Precision*	*Recall*	*F- score*
-2	0.09	0.45	0.15	0.83	0.24	0.35	0.08	0.45	0.13
-1	0.59	0.34	0.43	0.25	0.03	0.05	0.57	0.31	0.40
0	0.31	0.22	0.26	0.23	0.03	0.05	0.28	0.21	0.23
+1	0.65	0.52	0.57	0.43	0.94	0.59	0.63	0.43	0.51
+2	0.40	0.71	0.51	0.75	0.51	0.60	0.40	0.74	0.51
Accuracy	*0.436*			*0.443*			*0.395*		

To verify this, we reduced the size of the corpus (for all polarity classes) and run the experiment and found the accuracy values shown in Table 2.

Table 2: Relationship of Accuracy to Corpus Size

Size of Corpus	*Proportion*	*Accuracy*		
		Unigram	*Bigram*	*Hybrid*
310	50%	27.4%	36.5%	26.3%
458	75%	33.7%	43.5%	31.5%
608	100%	43.6%	44.3%	39.5%

It can be observed in Table 2 that the accuracy declines when the corpus size is reduced. This trend justifies that if the size of the corpus is increased, we expect an increment in accuracy which imply the model would perform much better. The percentage increment of accuracy in the bigram model, however, is relatively very low (less than 1% when the corpus size is increased from 75% to 100%). Since the accuracy of a classifier in bigram model is affected by the presence of valence shifters and our corpus has relatively small number of posts in the +2 and -2 classes which have higher concentration of valence shifters, the classifier may fail to capture as many bigram features as it does in the unigram feature even when the corpus size is increased.

5. Conclusion and Future Work

The enormous information available on the web has demanded mechanisms that categorize and present data in an understandable and usable format. Regarding this, sentiment analysis, which classifies texts based on polarity, is getting much attention recently.

In this paper, we presented a machine learning approach to multi-scale sentiment analysis on Amharic language. We used the Ethiopic script as it is by applying conversion to the Latin transliterated texts. For the machine learning task of sentiment analysis, we managed to collect around six hundred posts from online sources and applied lemmatization to tackle data sparseness problem which may arise due to morphological complexity of Amharic. To our knowledge, this is the first sentiment analysis approach on Amharic which applied morphological analysis and implemented using supervised machine learning approach.

It is our belief that we have rigorously followed the standard machine learning and evaluation approaches to sentiment analysis. Although a number of machine learning algorithms are available, we used Naïve Bayes algorithm to demonstrate the applicability of machine learning approach to Amharic sentiment analysis. This is due to its simplicity, efficiency of training and better performance in other morphologically rich languages. Our Naïve Bayes implementation to multi-scale sentiment analysis was successful and we achieved a promising performance accuracy of 43.6%, 44.3% and 39.5% for unigram, bigram and hybrid language models, respectively despite the few training data used. The results suggest the feasibility of multi-scale sentiment analysis as well as machine learning on Amharic. This has not been attempted by previous studies and it is essentially one of the major contributions of our work. We have also demonstrated that the accuracy can further be improved by building larger data set which was one of the most difficult and demanding tasks of our work. Our corpus is not only limited in size but also in diversity of contents as the majority of the posts in our collection are short, with no or limited number of emphasizing symbols, abbreviations, and short-hand writings. Moreover, Amharic dialectical terms are hardly available. Thus, we can also suggest that a rise in performance is also expected if a diversified training data is used. In addition, we are certain that morphological analysis has played its own role towards the encouraging results we have achieved. We were restricted to using only n-grams as a feature due to lack of NLP tools for Amharic that have limited our opportunities in using other features such as POS tagging.

Our work has significance on Amharic language research and usability in general and on sentiment analysis in particular. We strongly believe the

different components we have introduced in our sentiment analysis model have added their own values. The corpus and stop word list we have built as well as the different preprocessing tools we have developed have a potential to be used by future studies in related fields.

We have identified different research gaps that we believe could contribute to the sentiment analysis field on Amharic language. The first one is attempts on subjectivity analysis that deals with classification of texts as subjective and objective which can reduce the effort to be devoted in building sentiment corpus by avoiding manual identification of texts. The other direction of research can be document level sentiment analysis which is concerned with giving an overall document sentiment rather than at sentence and phrase level [3]. Finally, the concept-based approach to sentiment analysis which uses semantic knowledge bases to extract sentimental information within texts is assumed to perform better [11]. This has been attempted by relatively few researchers even for the developed languages such as English and should be given due consideration by future studies on Amharic.

References

[1] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79-86, 2002.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, Vol. 2, Issue 1-2, pp. 1-135, 2008.

[3] G. Vinodhini and R. M. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 6, 2012.

[4] J. Liu, M. Sarkar and G. Chakraborty, "Feature-based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio," SAS Global Forum 2013, 2013.

[5] B. Pang and L. Lee, "Seeing stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales," in Proceeding of the ACL, pp. 115-124, 2005.

[6] M. Abdul-Mageed, M. Diabc, and S. Kübler, "SAMAR: Subjectivity and Sentiment Analysis for Arabic Social Media," Computer Speech and Language, Vol. 28, No. 1, 2013.

[7] A. Mourad and K. Darwish, "Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs," in Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 55–64, 2013.

[8] S. Gebremeskel, "Sentiment Mining Model for Opinionated Amharic Texts," Unpublished Masters Thesis, Addis Ababa University, Addis Ababa, 2010.

[9] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment Analysis in the News", in Proceedings of the Seventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), 2010.

[10] M. Gasser, "HornMorpho: A System for Morphological Processing of Amharic, Oromo, and Tigrinya," in Proceeding of Conference on Human Language Technology for Development, 2011.

[11] E. Cambria, B. Schuller, and Y. Xia, "New Avenues in Opinion Mining and Sentiment," Intelligent Systems, IEEE, Vol. 28, Issue 2, pp. 15-21, 2013.