Sentence Level Amharic Text Sentiment Analysis Model: A Combined Approach

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

This paper propose the development of Amharic Sentiment Analysis model, whose goal is to classify the polarity of a piece of Amharic text according to the opinion of the writer. Since Amharic is one of under resourced language, the availability of tools to predict the sentiment of a sentence is very limited. There have been few proposals to alleviate this issue with, supervised machine learning and lexicon based approach. In this work, we combined these two approaches in order to give an account of their relative strengths and weaknesses. When run on a set of movie reviews the preliminary results provide interesting outcomes and pave the way for future research in the area. Our approach is a breakthrough in Amharic sentiment analysis, and opens exciting opportunities for future research.

1 Introduction

Today, methods for automatic opinion mining on online data are becoming increasingly relevant. Over the past few years, methods have been developed that can successfully and with a great degree of accuracy analyze the sentiment in opinions from digital English text.

Amharic is a Semitic language used for countrywide communication in Ethiopia. It is highly inflectional and quite dialectically diversified. With more than 20 million speakers, it is the second most spoken Semitic language in the World after Arabic (Asker, 2009). In spite of the relatively large number of speakers, it is still a language for which very few computational linguistic resources have been developed, and little has been done in terms of making useful higher-level Internet or computer-based applications available to those who only speak Amharic.

More and more digital information is now being produced in Ethiopia, but no deep-rooted culture of information exchange and dissemination has been established. Different factors are attributed to this, including lack of digital library facilities and central resource sites, inadequate resources for electronic publication of journals and books, and poor documentation and archive collections. Hence, this study aims to apply a combined technique for sentiment mining in order to discover hidden knowledge or pattern from an opinionated Amharic text.

2 Methods

The solution strategy adopted is described as follows.

2.1 Designing the model

Modeling is done by adopting the architecture of previously conducted research work (Gebremeskel, 2010) on the language and combining it with machine learning. Typically, lexicon-based approaches for sentiment classification are based on the insight that the polarity of a piece of text can be obtained on the ground of the polarity of the words which compose it. However, recent work in the area showed that supervised approaches tend to overcome unsupervised ones (Preslav, 2013; Preslav, 2014), the latter have the advantage of avoiding the hard-working step of labeling training data. Thus, in this model, the lexicon based method is first used for sentence orientation classification. Then after, by taking the output of the first classification method as an input, a machine learning classification model is designed

to classify the sentences in to positive and negative category. The machine learning classification module is designed after a detail investigation and understanding of the works by (Bing Liu, 2009; Bing Liu, 2012; Boiy and Moens, 2009).

2.2 Tools and Algorithms

Comments are extracted automatically from the web with the help of NLTK library. Python is used to implement the lexicon based algorithm. Before experimentation, data pre-processing task is performed to make the input reviews suitable for the NLTK tool. The translation is done using file conversion program in python. We have transliterated all texts into SERA. We selected support vector machines algorithm for supervised machine learning. WEKA is used to train and evaluate the performance of the machine learning classifier.

3 Experimental Evaluation

The lack of ready made available resources such as widespread lexical resources, data sources and well defined tools made conducting the experiment challenging. Table below provides comparison between average performances of all the three approaches in terms of overall accuracy, precision, recall and F-measure for the experimentation.

Approach	Average Recall	Average Precision	Average F Measure	Accuracy(%)
Lexicon Based	0.97	0.88	0.92	85%
Machine Learning	0.77	0.86	0.81	81.5%
Combined Approach	0.78	0.86	0.81	82.5%

Table 1: Experimental result.

4 Results and Discussion

In table 3.1 the best of all three approaches are presented with both F1 score and accuracy. The hybrid approach performs better than the machine learning approach, with a roughly 1% improvement in accuracy, while performing nearly 3.5% worse than the lexicon-based approach. The lexicon based technique has shown a better performance from the other approaches. One reason for this is the convergence of the sentiment terms used by reviewers and the collected Amharic sentiment terms in the lexicon. This is because the reviews collected are short in length and simple. If the length and complexity of reviews increases, the approach classification performance will decrease.

The proposed approach has achieved a better recall and precision than using the machine learning approach alone. The proposed method is able to quickly compute the sentiments of huge data sets with promising accuracy. Comments were analyzed with an accuracy of 82.5%.

5 Conclusions

The study has shown that sentiment analysis can be done automatically for Amharic texts by combining lexicon-based technique with machine learning without the need of having annotated corpus. As future work, we will extend the analysis by evaluating more lexical resources as well as more datasets. Moreover, we will refine our technique for threshold learning and we will try to improve our algorithm by modeling more complex syntactic structures as well as by introducing a word-sense disambiguation strategy to make our approach semantics-aware. Additional features that can be added to increase the performance of the proposed model should be studied in detail.

References

Asker L. Aragaw. Gamback B. Asfeha. and Habte 2009. *Classifying Amharic Webnews, in journal of information retrieval.*

Selama G. 2010. *Sentiment Mining Model for Opinionated Amharic texts*. M.S Thesis, Addis Ababa University, Addis Ababa.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. *Semeval-2013 task 2: Sentiment analysis in twitter*.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. *Semeval-2014 task 9: Sentiment analysis in twitter. Proc. SemEval*.

Bing Liu. 2012. *Sentiment Analysis and Opinion mining*. Morgan and Claypool publishers.

Bing Liu. 2009. *Sentiment Analysis, 5th Text Analytics Summit*. Boston.

Boiy E, Moens M. 2009. *A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts, information retrieval*. Cambridge University Press, Cambridge, UK.