Negation handling for Amharic sentiment classification

Girma Neshir[1], Andreas Rauber[2] and Solomon Atnafu[3]

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

Introduction: Due to the advancement of World Wide Web technology, users usually express their feelings, emotions and opinions as comments in response to the posted news, photo, audio and video. Currently, opinionated sources are increasing in languages other than English. However, Amharic sentiment analysis researches are very few as it has no sufficient linguistic resources for linguistic preprocessing and sentiment analysis. There are several challenges in lexicon based sentiment analysis. One of these is that handling negation in the text. The most common approach for negation handling is carried out relying on negation keywords. However, it is complex to identify the scope of negation where the process of correctly identifying the part of the text affected by the presence of negation word. Negation Handling(NH) is never studied in Amharic language to the best of our knowledge. Thus, this research develops an automatic method to handle negation and combined with char ngram features for Amharic sentiment classification. The research questions to be addressed in this work are as follows: (a) how can we automatically detect negation words in Amharic texts? (b) how can we design a framework for handling negation in Amharic sentiment analysis? (c) how to capture char level ngram features for improving Amharic sentiment analysis in Social media(e..g. Facebook) and (d) how can we evaluate the performance of the framework?

Proposed Approach: As part of preprocessing, we normalized not only all Amharic words in the Amharic News Comments but also handling entries of Amharic Sentiment Lexicon by replacing varied alphabets of the same sound with identical symbols. Moreover, a stemmer is applied after negation identification is completed. As Amharic is morphologically rich, to reduce the mismatch of Amharic words during string comparison operation. We also used stemming for this purpose. The proposed framework consists of components including preprocessing and sentiment score calculation using negation detection and machine learning using char level ngrams features. For more detail, the proposed framework is shown in Fig. 1 of Appendix. To compute sentiment score using negation detection, for each Amharic news comment, C_i, if each stemmed word w_{ij} is found in either of the Amharic Sentiment lexicons (Manual, SOCAL, SWN) [7], then the sentiment score s_{ij} is retrieved. s_{ij} and its position index in the comment is stored. To compute the sentiment of the comment, we apply positional weighting inversion if the comment contains any negation clue. If negation clue is not found, the score of the word is simply added. For more detail, the negation handling algorithm is depicted in Listing 1 of Appendix. Besides lexicon based negation handling approach, the usefulness of character aware language models is well suited to apply for language identification, reducing of text feature sparse di-

mensionality, helps to handle spelling errors, abbreviations, special characters, etc. [6]. That is why we proposed character level ngram approaches to reduce and address these issues for Amharic facebook news comments' sentiment classification rather than word level ngram approaches. For example, the negation carrying Amharic word "አልወደውም"/ "I do not like him"/ has 2-gram character level features includes አ-አል-ልወ-ወደ-ደው-ው-ም-ም, 3-gram character level features are አ- አል-አልወ-ልወደ-ወደው-ደው-ም-ው-ም-ም, etc. The negation marker/morpheme/ አል- is detected as feature of the negation word/"አልወደውም"/. Thus, prior to char ngram based sentiment classification, we partition the Annotated Amharic Facebook News Comment corpus into training and testing sets. Logistic regression(LR) and Naïve Bayesian(NB) models are built relying on the term-frequency inverse document frequency(tfidf) of char level and word level(baseline) bi-gram and tri-gram features of training set.

Results: For Amharic Sentiment Classification, the results of the accuracy of the individual and the combined models on the test set are presented in detail in Table 1 of Appendix. The results in Table 1 show that negation handling algorithm outperforms very well (acc. 86.2%) than the performance of character level and word level based machine learning models for classifying sentiment of Amharic texts. On the other hand, character level ngram based classifier is more useful for classifying Amharic Sentiment than word level ngram models (baseline) (accuracy of 95.27). Finally, the hybrid model is obtained by combining negation handling approach and char ngram models (NB+LR). This hybrid model outperforms with accuracy of 98% than the other models and its combinations. Yet, it is quite difficult to find why errors are generated in predicting sentiment category of Amharic Facebook Comments. For example the facebook comment, 'በቃል የሚነገሩነገሮችን በተግባር እንዲፈፀሙልን እንፈልጋለን፣ '/ We need to see accomplished in practice that we heard in words/ is wrongly predicted. This comment does not express any opinion. This kind of comment represents wishes that someone wants it to be done, but not necessarily expressing sentiment. Further researches needs to be carried out to reduce the source of errors in predicting sentiment class of Amharic comments. Our recent findings is a good starting point to improve the performance of Amharic sentiment analysis in facebook news comments. Fine tuning char ngram features shows suitableness and flexible for sentiment analysis of resource limited language (e.g. Amharic) than word level ngram models

Conclusions: In general, extensive linguistic resources are expensive to build sentiment classification on the less dominant languages (e.g. Amharic). To reduce this problem, we proposed negation handling approach and char ngram approach for Sentiment analysis of Amharic face book news comments. So far, we have seen that the proposed approach still lacks accuracy of Amharic sentiment classification. The approach potentially does not sufficiently capture the language specific features that help to identify the sentiment class of Amharic news comment in social media. Further work should be performed to reduce the amount of errors in sentiment analysis of Amharic facebook news comments. To address these issues, we may need to consider char ngram embedding features from corpus of the same domain(e.g. Facebook news comments). Besides, Amharic negation scope identification and handling is recommended to be investigated for further researches.

References

1. Rizkiana Amalia, Moch Arif Bijaksana, and Dhinta Darmantoro. Negation handling in sentiment classification using rule-based adapted from indonesian language syntactic for indonesian text in twitter. In Journal of Physics: Conference Series, volume 971, page 012039. IOP Publishing, 2018.
2. Amna Asmi and Tanko Ishaya. Negation identification and calculation in sentiment analysis. In The second international conference on advances in information mining and management, pages 1--7, 2012.
3. Claudia Diamantini, Alex Mircoli, and Domenico Potena. A negation handling technique for sentiment analysis. In 2016 International Conference on Collaboration Technologies and Systems (CTS), pages 188--195. IEEE, 2016.
4. Martine Enger, Erik Velldal, and Lilja Øvrelid. An open-source tool for negation detection: a maximum-margin approach. In Proceedings of the Workshop Computational Semantics Beyond Events and Roles, pages 64--69, 2017.
5. Umar Farooq, Hasan Mansoor, Antoine Nongaillard, Yacine Ouzrout, and Muhammad Abdul Qadir. Negation handling in sentiment analysis at sentence level. JCP, 12(5):470--478, 2017.
6. Bas Heerschop, Paul van Iterson, Alexander Hogenboom, Flavius Frasincar, and Uzay Kaymak. Accounting for negation in sentiment analysis. In 11th DutchBelgian Information Retrieval Workshop (DIR 2011), pages 38--39, 2011.
7. Girma Neshir Alemneh, Andreas Rauber, and Solomon Atnafu. Dictionary Based Amharic Sentiment Lexicon Generation, pages 311--326. 08 2019.

Appendix: List of figures, tables and algorithm listings

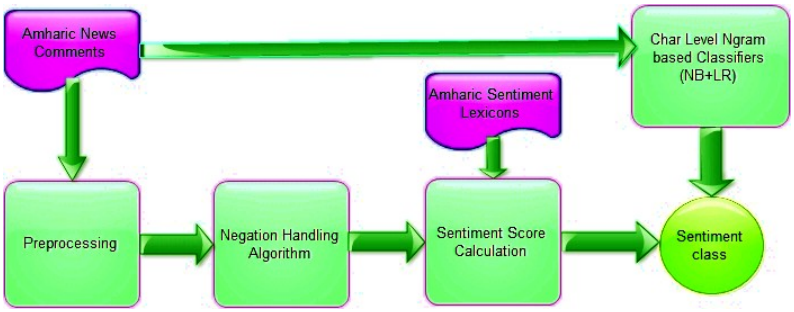

Fig. 1 Proposed Framework for Negation Handling

DetectNegation(W_{ij}):
Input= The jth *Amharic word* W_{ij} *of the* i^{th} *comment.*
NegationList= Load Amharic Negation list from file;
Prefix List=initialize by Amharic negation prefixes
Suffix List= initialize by Amharic negation suffixes

If W_{ij} *begins with one of the PrefixList and len*$(W_{ij})>1$*:*
    Return True
If W_{ij} *ends with one of the SuffixList and len*$(W_{ij})>1$*:*
    Return True
If W_{ij} *match with one of the words in NegationList:*
    Return True
Else:
    Return False

Listing 1. Snap shot Algorithm for Amharic Negation Handling

Table 1. depicts the classification performance of the proposed framework

Approach	**Accuracy**	
	Word Level (baseline)	Char Level
Negation Handling(NH)	86.2	-
NB + LR	79.32	83.75
Hybrid(NH+NB+LR)	**95.27**	**98.0**