Sentiment Analysis amidst Ambiguities in YouTube Comments on Yoruba Language (Nollywood) Movies

Orimaye Sylvester Olubolu
Faculty of Information Technology
MONASH University, Malaysia
sylvester.orimaye@monash.edu

Saadat M. Alhashmi
Faculty of Information Technology
MONASH University, Malaysia
alhashmi@monash.edu

Siew Eu-gene
Faculty of Information Technology
MONASH University, Malaysia
siew.eu-gene@monash.edu

ABSTRACT
Nollywood is the second largest movie industry in the world in terms of annual movie production. A dominant number of the movies are in Yoruba language spoken by over 20 million people across the globe. The number of Yoruba language movies uploaded to YouTube and their corresponding comments is growing exponentially. However, YouTube comments made by native speakers on Yoruba movies combine English language, Yoruba language, and other commonly used "pidgin" Yoruba language words. Since Yoruba is still a resource constrained language, existing sentiment or subjectivity analysis algorithms have poor performances on YouTube comments made on Yoruba language movies. This is because of the constrained language ambiguities. In this work, we present an automatic sentiment analysis algorithm for YouTube comments on Yoruba language movies. The algorithm uses SentiWordNet thesaurus and a lexicon of commonly used Yoruba language sentiment words and phrases. In terms of precision-recall, the algorithm performs more than a state-of-the-art sentiment analysis technique by up to 20%.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]: Text Analysis—*Sentiment Analysis*

General Terms
Algorithms, Experimentation

Keywords
Sentiment analysis, natural language application, Yoruba lexicon

1. INTRODUCTION
Sentiment analysis on movie comments has become very useful in improving the quality of movies and satisfying the demands of the consumers. The number of movies and comments on YouTube is growing exponentially [4]. With sentiment analysis, substantial comments can be mined for *positive*, *negative* and *neutral* sentiments. This can help in satisfying viewers' experience and automatic recommendation of movies to interest groups. However, sentiments expressed using a resource constrained language pose very difficult task for existing sentiment analysis algorithms. In this work, we perform sentiment analysis for YouTube comments on Yoruba language movies. Yoruba language is resource constrained and the native speakers often use a combination of the language, English language, and other "pidgin" Yoruba language words in YouTube comments. This causes ambiguity problem for existing sentiment analysis algorithms and therefore affects their efficiencies. For example, consider the following sentence from a YouTube comment on a Yoruba movie.

WWW 2012 Companion, April 16-20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

(1) *Yes oooooo! I gbádùn dis movie láti òkè dé ìsàlè.*

In the sentence above some words that express sentiments are actually written in Yoruba language or "pidgin" Yoruba language. A formalized interpretation of the above sentence in English language is as follows:

(2) *Great!* I *enjoyed* this movie from the beginning to the end.

From this interpretation, words that express sentiments are shown in italics. In the context of YouTube comments on nollywood movies, the phrase *Yes oooooo!* is a "pidgin" expression of *great* or *interesting mood*. The frequency of alphabet 'o' in the phrase can also vary depending on the intensity of the mood. For example, *yes ooo, yes oooooo*, and *yes ooooooooo*. Whereas, such words are seen as ambiguities by the existing sentiment analysis algorithms. While it could be tempting to rely on the number of "likes" and "dislikes" in YouTube comments to measure the overall sentiments on a movie, the two numbers combined is often lower than the total number of comments and thus cannot show the absolute sentiment distribution on the movie.

In this paper, our contributions are twofolds. First, we develop a sentiment lexicon for Yoruba language. The lexicon contains Yoruba language words and phrases commonly used to express sentiments in YouTube comments on Yoruba movies. We manually annotate such words based on their polarity of sentiments. We evaluate the annotations using Inter-Annotator Agreement Study. Finally, we develop a sentiment analysis algorithm that interpolates the score of English sentiment words with the score of Yoruba sentiment words to give a final sentiment polarity score for each comment.

2. SENTIMENT ANNOTATION

2.1 Corpus
For the purpose of this work, a corpus containing YouTube comments on Yoruba language movies has been created. The corpus contains 15,000 YouTube comments manually collected from 270 Yoruba language movies between May, 2011 and July, 2011. The detail of the corpus is shown in Table 1. The corpus and our sentiment analysis algorithm is also publicly available for research purpose[1].

Table 1. Details for corpus of YouTube comments on Yoruba movies

A corpus of YouTube comments on Yoruba Language Movies	
Total number of comments in the corpus	15,000
Total number of sentences in the corpus	39,231
Average number of sentences in a comment	2.5
Average number of distinct words in a comment	15

2.2 Annotation
Using the corpus above, we asked participants to identify and label Yoruba words, phrases, and "pidgin" yoruba words that express "positive" or "negative" or "neutral" sentiments in each comment.

[1] https://sourceforge.net/projects/yorubasentiment/

We do not take "mixed" sentiments into account since the study conducted in [2] reported that "mixed" category gives low agreement at minimal granulity such as sentences found in YouTube comments. We then measure inter-annotator reliability for all comments in the corpus. Our 5 participants were postgraduate students whom are all native Yoruba speakers including one whom has worked on sentiment analysis. In the annotation process, participants were asked to score each sentiment word or phrase labeled with a value between 0 and 1. In total 1,076 words and phrase were annotated out of which 478 are positive, 301 are negative, and 297 are neutral.

2.3 Inter-Annotator Agreement

We used the statistical method adopted in [2] to measure our inter-annotator agreement study. This is beacause our corpus is large and the statistical method does not require that each annotator annotates each comment interchangeably as the observed variability is due to chance. The α value gotten as a result of the statistical method is 0.624 for the 3 sentiment categories highlighted earlier. This value indicates substantial agreement between annotators as it is close to the recommended *tentative reliability* value of 0.667.

3. THE PROPOSED ALGORITHM

The algorithm does a *linear combination* of the sentiment score of English "adjectives" and "adverbs" derived from SentiWordNet thesarus [1] and the sentiment score of Yoruba words derived from our Yoruba sentiment lexicon (Section 2). We selected English adjectives and adverbs because they are naturally used to express sentiments (e.g. *very interesting movie*). Because we need to sum the sentiment scores for English and Yoruba words, we compute the sentiment scores for English and Yoruba words independently. Both scores are then linearly combined to determine the leading sentiment in a comment. Thus, there are two steps involved.

Step 1: Polarity score for each English word is determined using SentiWordNet. SentiWordNet returns the polarity scores of corresponding synsets for each English word but returns "null" for non-English words. The polarity scores may be different for each of the synset entries. The absolute polarity scores for all the synset entries (i.e. positive, negative, neutral) is computed as follows:

$$Score_\lambda(w) = \frac{1}{k}\sum_{i=1}^{k}\tau_{pos}(e_i)\ , \quad \frac{1}{k}\sum_{i=1}^{k}\tau_{neg}(e_i), \quad \frac{1}{k}\sum_{i=1}^{k}\tau_{neu}(e_i) \quad (1)$$

where $Score_\lambda(w)$ is the absolute score for each polarity of the word w based on the number of synset entries k and τ is the SentiWordNet polarity score for each synset entry e. Thus, the $Score_\lambda(w)$ for "each" word would give 3 absolute polarity scores for all its synsets (i.e. positive, negative, and neutral). A naïve approach to compute the sentiment score for English words is to find and sum the maximum polarity score from each word synset. However, this approach would give a constant polarity score for every word in SentiWordNet since k is constant. This ignores the importance of the different synsets polarity scores and counterintuitive to the overall sentiment distribution in a comment. We compute the sentiment score for all English words as follows:

$$Score_\lambda(E_c) = \max_{Score}\left\{\sum_{i=1} Score_{pos}(w_i),\ \sum_{i=1} Score_{neg}(w_i), \sum_{i=1} Score_{neu}(w_i)\right\} \quad (2)$$

where $Score_\lambda(E_c)$ is the maximum sentiment score for all English words in the comment, $Score_{pos}(w_i)$, $Score_{neg}(w_i)$, $Score_{neu}(w_i)$ are the absolute sum of positive, negative, and neutral synsets scores respectively.

Step 2: We use *n*-gram technique to pick words from the comment and compare with words in our Yoruba sentiment lexicon. Because the lexicon also contains Yoruba phrases, the *n*-gram technique seems appropriate for our work. We set our n to 4 as the longest phrase in our lexicon contains 4 words. If our lexicon contains the *n*-gram selection, it returns the polarity score. Like the English words, we compute the absolute polarity scores for Yoruba sentiment words or phrases by summing all positive, negative, and neutral scores if any, otherwise each polarity score is set to "zero". However, we do not consider Yoruba synsets like SentiWordNet as it is beyond the scope of this study. We select the maximum polarity score as the absolute sentiment score for Yoruba words as follows:

$$Score_\lambda(Y_c) = \max_{Score}\left\{\sum_{i=1} Score_{pos}(y_i),\ \sum_{i=1} Score_{neg}(y_i), \sum_{i=1} Score_{neu}(y_i)\right\} \quad (3)$$

where $Score_\lambda(Y_c)$ is the maximum sentiment score for all Yoruba words in the comment, $Score_{pos}(y_i)$, $Score_{neg}(y_i)$, $Score_{neu}(y_i)$ are the absolute sum of positive, negative, and neutral polarity scores respectively. Finally, the absolute sentiment score for each comment is computed as the linear combination of equation 2 and 3 satisfying a linear expression $C = \alpha X + \beta Y$.

$$Score_\lambda(E_cY_c) = \frac{1}{\alpha}\,Score_\lambda(E_c) + \frac{1}{\beta}Score_\lambda(Y_c) \quad (4)$$

where α and β are the number of English and Yoruba words in each comment respectively.

4. EXPERIMENTS AND RESULTS

We selected 10,000 comments at random from our dataset and performed experiment with our algorithm and another state-of-the-art algorithm [3]. We then compared the results of both algorithms in terms of precision and recall. We selected [3] because it uses only SentiWordNet for multilingual sentiment analysis (excluding Yoruba). Table 2 shows the precision and recall values for both algorithms and Figure 1 shows the precision and recall curves for random 10 comments (assuming a movie with 10 comments).

	Precison		
	Pos	Neg	Neu
Ours	85%	87%	79%
[3]	61%	64%	67%
	Recall		
Ours	91%	89%	90%
[3]	68%	56%	65%

Table 2: Precision-Recall results

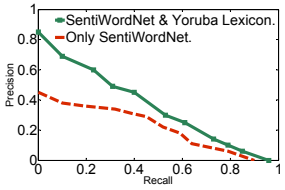

Figure1: Precision-Recall curves

The purpose of figure 1 is to show distinct difference in performances between the two algorithms at a very minimal level. It is remarkable that our algorithm shows significant performance at both large scale and minimal level. Even if Yoruba language is not resource constrained, we think our algorithm will still perform better because of its ability to account for "pidgin" Yoruba words.

5. CONCLUSION

We proposed automatic sentiment analysis algorithm that performs better on YouTube comments on Yoruba movies using SentiWordNet thesaurus and Yoruba sentiment lexicon. We will extend our work to detect Yoruba sentiments from Facebook updates for the purpose of placing contextual advertisements.

6. REFRENCES

[1] Esuli, A., Sebastiani, F. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, LREC, 2006.
[2] Bermingham, A. and Smeaton, A.F. A study of inter-annotator agreement for opinion retrieval, ACM SIGIR, 2009.
[3] Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis, IEEE ICDEW 2008.
[4] Siersdorfer, S., Chelaru, S., Pedro, J.S. How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings, WWW 2010.