Semantic Enrichment of Nigerian Pidgin English for Contextual Sentiment Classification

Wuraola Fisayo Oyewusi[1*] , **Olubayo Adekanmbi**[2] , **Olalekan Akinsande**[3]

[1]Data Science Nigeria
[2]Data Science Nigeria
[3]Data Science Nigeria
{wuraola, olubayo, olalekan}@datasciencenigeria.com,

Abstract

Nigerian English adaptation, Pidgin, has evolved over the years through multi-language code switching, code mixing and linguistic adaptation. While Pidgin preserves many of the words in the normal English language corpus, both in spelling and pronunciation, the fundamental meaning of these words have changed significantly. For example, 'ginger' is not a plant but an expression of motivation and 'tank' is not a container but an expression of gratitude. The implication is that the current approach of using direct English sentiment analysis of social media text from Nigeria is sub-optimal, as it will not be able to capture the semantic variation and contextual evolution in the contemporary meaning of these words. In practice, while many words in Nigerian Pidgin adaptation are the same as the standard English, the full English language based sentiment analysis models are not designed to capture the full intent of the Nigerian pidgin when used alone or code-mixed. By augmenting scarce human labelled code-changed text with ample synthetic code-reformatted text and meaning, we achieve significant improvements in sentiment scoring. Our research explores how to understand sentiment in an intrasentential code mixing and switching context where there has been significant word localization.This work presents a 300 VADER lexicon compatible Nigerian Pidgin sentiment tokens and their scores and a 14,000 gold standard Nigerian Pidgin tweets and their sentiments labels.

1   Introduction

Language is evolving with the flattening world order and the pervasiveness of the social media in fusing culture and bridging relationships at a click. One of the consequences of the conversational evolution is the intrasentential code switching, a language alternation in a single discourse between two languages, where the switching occurs within a sentence [Koban, 2013]. The increased instances of these often lead to changes in the lexical and grammatical context of the language, which are largely motivated by situational and stylistic factors [Inuwa *et al.*, 2014]. In addition, the need to communicate effectively to different social classes have further orchestrated this shift in language meaning over a long period of time to serve socio-linguistic functions [Ifechelobi, 2015] Nigeria is estimated to have between three and five million people, who primarily use Pidgin in their day-to-day interactions. But it is said to be a second language to a much higher number of up to 75 million people in Nigeria alone, about half the population.[Carons and Onyioha, 2012]. It has evolved in meaning compared to Standard English due to intertextuality, the shaping of a text's meaning by another text based on the interconnection and influence of the audience's interpretation of a text. One of the biggest social catalysts is the emerging urban youth subculture and the new growing semi-literate lower class in a chaotic medley of a converging megacity [Igboanusi, 2008] [Samanta *et al.*, 2019] VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains. VADER lexicon has about 9000 tokens (built from existing well-established sentiment wordbanks (LIWC, ANEW, and GI) incorporated with a full list of Western-style emoticons, sentiment-related acronyms and initialisms (e.g., LOL and WTF)commonly used slang with sentiment value (e.g., nah, meh and giggly)) with their mean sentiment rating.[Hutto and Gilbert, 2014]. Sentiment analysis in code-mixed text has been established in literature both at word and sub-word levels [Prabhu *et al.*, 2016] [Roncal, 2019] [Jang and Shin, 2010]. The possibility of improving sentiment detection via label transfer from monolingual to synthetic code-switched text has been well executed with significant improvements in sentiment labelling accuracy (1.5%, 5.11%, 7.20%) for three different language pairs [Samanta *et al.*, 2019]

2   Method

This study uses the original and updated VADER (Valence Aware Dictionary and Sentiment Reasoner) to calculate the compound sentiment scores for about 14,000 Nigerian Pidgin

---

*Contact Author

Pidgin Words	VADER Token and Score	Avg. Score
kasala	riot(-2.6), trouble(-1.7)	-2.2
gbege	catastrophe (3.4), chaotic(-2.2)	-2.9
para	angry(-2.3), rage(-2.6)	-2.2

Table 1: Average Sentiment Score for Nigerian Pidgin Sentiments with Multiple English Meanings

tweets[1]. The updated VADER lexicon (updated with 300 Pidgin tokens[2] and their sentiment scores) performed better than the original VADER lexicon. The labelled sentiments from the updated VADER were then compared with sentiment labels by expert Pidgin English speakers.
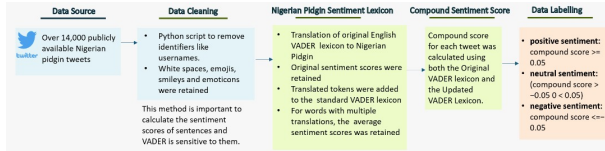

Figure 1: The semantic enrichment of Nigerian pidgin English for contextual sentiment classification methodology.

3 Results

During the translation of VADER English lexicon to suitable one-word Nigerian Pidgin translation, a total of 300 Nigerian pidgin tokens were successfully translated from the standard VADER English lexicon. One of the challenges of this translation is that the direct translation of most the sentiment words in the original VADER English Lexicon translates to phrases not single one-word tokens and certain pidgin words translates to many english words.**??**.

A Appendix

A.1 Selection of Data Labellers

Three people who are indigenes or lived in the South South part of Nigeria, where Nigerian Pidgin is a prevalent method of communication were briefed on the fundamentals of word sentiments. Each labelled Data point was verified by at least one other person after initial labelling.

References

[Carons and Onyioha, 2012] Tosin Carons, Abraham and M Amaka Onyioha. The origin of pidgin. *Afrostyle Magazine*, 2, 2012.

[Hutto and Gilbert, 2014] C.J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, 2014.

[Ifechelobi, 2015] Jane Ifechelobi. Code switching: a variation in language use. *Mgbakoigba: Journal of African Studies*, 4:1–7, 2015.

[Igboanusi, 2008] Herbert Igboanusi. Empowering nigerian pidgin: A challenge for status planning? *World Englishes*, 27:68 – 82, 02 2008.

[Inuwa *et al.*, 2014] Yusuf Inuwa, Nuhu, Anne Christopher, Althea, and Haryati Bakrin, Bt. Factors motivating code switching within the social contact of hausa bilinguals. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*, 19:43 – 49, 2014.

[Jang and Shin, 2010] Hayeon Jang and Hyopil Shin. Language-specific sentiment analysis in morphologically rich languages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 498–506. Association for Computational Linguistics, 2010.

[Koban, 2013] Didem Koban. Intra-sentential and inter-sentential code-switching in turkish-english bilinguals in new york city, u.s. *Procedia - Social and Behavioral Sciences*, 70:1174–1179, 01 2013.

[Prabhu *et al.*, 2016] Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. 11 2016.

[Roncal, 2019] Iñaki San Vicente Roncal. *Multilingual sentiment analysis in social media*. PhD thesis, Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2019.

[Samanta *et al.*, 2019] Bidisha Samanta, Niloy Ganguly, and Soumen Chakrabarti. Improved sentiment detection via label transfer from monolingual to synthetic code-switched text. *arXiv preprint arXiv:1906.05725*, 2019.

---

[1]Link to Nigerian Pidgin tweets and Sentiments https://git.io/JvHrp.

[2]Link to 300 Nigerian Pidgin Sentiments and Scores https://git.io/Jv9og.