Department: Affective Computing and Sentiment Analysis
Editor: Erik Cambria, Nanyang Technological University

This! Identifying new sentiment slang through orthographic pleonasm online: Yasss slay gorg queen ilysm

Mike Thelwall
University of Wolverhampton

***Abstract*—Identifying neologisms is important for natural language processing of social web text when informal language is standard and youth slang is common. For example, failing to identify neologisms can reduce the accuracy of lexical sentiment analysis if opinions are frequently expressed in words that are too new to be in the sentiment dictionary. This article proposes a method based on orthographic pleonasm to identify emotion-related neologisms in the social web: finding words with the most different letter repetition spelling variations. For this method, non-dictionary words are extracted from a large social web corpus, spelling standardisation is applied, and then words are ranked in decreasing order of spelling variation frequency. Words with the most spelling variations are then KWIC-analysed for semantic context. Applied to a collection of comments on YouTube influencers, this method found neologisms like *slay* and *early* as positive terms, mixed with traditional sentiment words, exclamations, and nouns. Although orthographic pleonasm was originally used to express the speaker's rhythm and one of voice, it is also used for initialisms in a way that is difficult to vocalise. The method is therefore a practical method to identify new sentiment slang, including both normal words and initialisms.**

INTRODUCTION Natural language processing (NLP) algorithms applied to social web texts can struggle with non-standard spellings and grammar, especially if they are optimised for formal documents or applied to texts that are rich in nonstandard features (e.g., [1]). Slang words are a problem because of this and because they vary between subgroups in society and evolve over time [2]. Since new slang words periodically appear and are extensively used by young people, NLP algorithms need to be able to be updated for these terms to perform optimally. For example, sentiment analysis algorithms [3] may need to exploit slang [4] and may be fooled when new sentiment slang terms become the norm amongst some social groups [5], such as (in the past) *dad*,

 This article is part of the IEEE IS ACSA series

lush skill, *tidy*, and *wicked*. Thus, a method is needed to identify new sentiment slang. Although slang is sometimes manually identified [6], an automatic or semi-automatic method might be more efficient. In practical terms is possible to consult slang dictionaries [7], [8], [9], [5], but these can be large, of uncertain provenance and general purpose rather than tailored for sentiment. Similarly, slang-based extensions to linguistic resources like WordNet [10] may become dated.

Lexicographic techniques to identify new words rely on frequency to find them, such as by identifying the most common non-dictionary words found in a social web corpus. Whilst this works for neologisms like *lol*, it will not work for existing words repurposed as slang, including the examples above. Topic modelling can be used to check if a word has changed meaning, potentially with a new slang meaning, relative to changes in discussion topics (experiment 2 of: [11]). This method does not identify new words, however, and is not specific to sentiment. Deep learning has been used to identify dictionary sentences with slang from Wall Street news sentences, finding that slang commonly associates with syntactic changes or shifts [12]. This method is promising but may not work well on informally structured language. Slang words not in a dictionary can also be identified through web context information, but this does not work for slang meanings of standard words [13]. Sentiment words can be also extracted from dialect language, including both slang and standard terms [14], although this is a limited context. With a similarly restricted context, polarised creative political slang can be found by comparing word frequencies between sets of comments from different political positions, after removing dictionary words, proper nouns and known slang [15]. Rhyming slang has special detection algorithms [16] but is a rare type. Finally, a semi-automatic method has also been devised to create an annotated corpus of youth slang [17]. In summary, whilst many methods exist to detect slang or, in limited contexts, sentiment slang, none detect new sentiment slang in a general context.

This article harnesses a relatively unique style of informal social web text to detect new sentiment terms. Computer-mediated communication encourages the development of styles that adapt to the medium. Informal conversational messages may be misinterpreted when written because they lack the visual cues that are necessary to interpret them. This has led to devices like emoticons to either inject sentiment or indicate the type of sentiment intended by the author [18]. The emoticon therefore substitutes for the ability to read the speaker's expression. A second device is to spell words in the way that they are spoken, injecting extra letters to indicate emphasis or enthusiasm, sometimes called "vocal spelling" [19], [20]. It is orthographic pleonasm through the inclusion of (relatively) redundant letters. This substitutes for the ability to interpret a person's tone of voice or intonation. It can also be used to inject a regional accent into words [21]. Both emoticons and spelling alterations seem to be widely enough used to be easily decoded by the reader. Capitalisation, underscores, asterisks and brackets can also be used for emphasis or to suggest increased volume of speech [22]. Whilst emoticons and altered spelling have been used in sentiment analysis as indicators of sentiment [23], [24], they have not been used to detect slang words.

This article proposes a new method to identify new sentiment slang from a set of social web texts by exploiting orthographic pleonasm: redundant letters added to a word. For example, someone might comment wickiiiiiiid on Facebook, with the redundant letters functioning to emphasise the sentiment of the word. The intuitions behind using orthographic pleonasm to identify new sentiment-related words are that (a) slang spelling is likely to associate with slang words because both are types of informal language, so slang words seem likely to be written with extensive orthographic pleonasm, and (b) emotional states can lead to changed tones of voice, so vocal spelling can be expected for sentiment words. Although there are other creative spelling techniques than repeating letters, such as leet (e.g., [25]), orthographic pleonasm seems to have the most natural association with sentiment.

The orthographic pleonasm semi-automatic method to detect sentiment slang has four stages and requires a collection of contemporary social web texts.

1) Extract all words from a set of social web

texts without any spell checking.
2) Convert each word into a common standard form by deleting repeated letters until a dictionary word is found or all doubled letters have been removed (e.g., lushhhh to lush). This is the shortened form of the word.
3) Count the number of spelling variations of each shortened word form, and rank the short word forms in descending order of this number.
4) Manually check the top N words for meaning using the Key Word In Context KWIC method to identify the sentiment slang.

EXPERIMENT

The orthographic pleonasm method was applied to a set of nine million YouTube comments on the videos of 223 UK lifestyle-based female influencers. These were downloaded in November-December 2020 with the YouTube API through the free software Mozdeh. A set from the same country as the author is helpful for slang identification and lifestyle influencer comments are particularly informal and sentiment rich, so are a suitable source. They have the additional practical benefit that they are easy to collect, seem to be relatively spam-free, and are date-stamped. Individual YouTube commenters can sometimes be prolific, however, which may skew the results towards the writing styles of a few people. To guard against this, each commenter was limited to one comment per month, using a random number generator (in Mozdeh). This is a compromise that gives a moderate amount of extra weight to prolific commenters. To further guard against domination by individuals, commenters most using each word were checked during the follow-up KWIC process (again, an option in Mozdeh), for being the single dominant users of the term. No causes for concern were discovered in this regard. This stage produced 13,264,344 comments from July 2007 to December 2020.

The orthographic pleonasm method was applied, as described above, and the top 100 terms were manually checked by the author for slang uses using the KWIC approach (using Mozdeh's concordancer) with the original YouTube comments. Each word was also classified for primary function to give supporting contextual information about orthographic pleonasm in practice.

RESULTS

Thirty-six of the 100 words examined were common sentiment terms, with the extra spelling serving to boost the strength of the sentiment expressed. There were many laugh variants (*hahahahahahaha, hahaha, hahahahahahaha, hahahahaha, hahahaha, hahahahahaha, haha*) and two abbreviations for laughter (*lmao, lol*). Whilst none of these are traditional sentiment terms, they seem to have a long online history. There were also positive opinion words (*love, amazing, beautiful, loved, cute, gorgeous, pretty, awesome, stunning, super, perfect, best, great, nice, congrats, lurve, crazy*). Several non-sentiment words expressed a positive sentiment in their typical contexts. These included (waited for) *long*, (waited for) *ages*, *finally* (it is ready), (can't) *wait*, (love) *it*, (want) *more*, (this is) *huge*. Two also expressed emotions: *happy, crying*. The latter word was expressed typically in positive contexts like weddings and after seeing a new pet or flat. Many different numbers of kisses (x) were also sent. The remaining words were relatively new ways of expressing sentiment (the primary target of this article), or terms performing different functions. These are discussed separately by function.

New sentiment slang

Eleven terms were new slang words (Table 1), two being related abbreviations (e.g., *ilysm*: I love you so much). One of the exclamations, *omg*, is a relatively old abbreviation and so is not included in this set. All these words were often used on their own in comments or with emoticons, emphasising that the word itself expresses the sentiment. For example, *goals* on its own apparently carried the meaning "excellent" .

The new slang words displayed different but overlapping patterns of use (Figure 1). This is complicated by some being used in non-slang contexts (e.g., early). Whilst several words seem to have fallen out of fashion (e.g., *ilysm, slay, slaying, goals, yas, dayum*), others were still fashionable at the time of data collection (e.g., *early, dying, queen*).

Other new slang was searched for less comprehensively within the words ranked 101-300 for spelling variations, finding many additional

Table 1. New slang terms found in the 100 words with the most spelling variations.

Word	Comments	Spellings
yas	40495	492
early	70732	196
goals	1077	168
slay	12109	155
ilysm	15980	124
dying	21808	123
slaying	4592	107
dayum	818	103
queen	687	103
lysm	8390	97

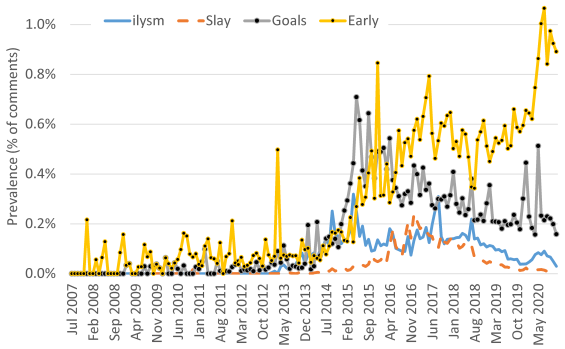

Figure 1. Percentage of comments containing each term, by month, up to November 2020.

terms: *this* (88 variants), *screaming* (84), *slayed* (76), *ahmazing* (74), *sick* (69), *bomb* (69), *preach* (66), *gorg* (55), *lush* (54, also a brand), *omfg* (53, quite old), *poppin* (50), and *lit* (49). The pronoun/adverb/determiner *this* was often used in a traditional context, but also as a single word determiner, presumably to endorse the content or message of the video.

Words not directly expressing sentiment

Some of the terms with the most spelling variations performed functions other than directly expressing sentiment. Seven of the top 100 terms were booster words or had a booster function (Table 2). These included standard booster words like *really*, *so* and *very*. The words *way* and *much* commonly occurred in phrases with a booster function, such as, "I love this way too much." Lengthening the spelling of these words served to emphasise (or boost) their booster function, a natural linguistic marriage.

Twenty terms were exclamations – usually used on their own, with emoticons, or to introduce a complement (Table 3). These are a natural fit

Table 2. Booster words found in the 100 words with the most spelling variations.

Word	Comments	Spellings
really	813147	413
so	3452010	291
very	355367	130
ever	219519	177
everything	178118	155
way	303933	138
much	822935	161

with additional letters since uttered exclamations are often drawn out and adding extra letters would mimic this.

The list includes standard exclamations like *wow* and *argh* as well as words that have other functions but were mainly used as exclamations. *God, damn*, and *fuck* were used as exclamations and swear words (e.g., fuuuck your hair is gorgeous). The terms here typically expressed excitement, but this excitement is probably interpreted as positivity in context since the comments tended to be positive overall. Within this list, *yay, wow*, and *woohoo* could also be classified as positive sentiment terms, such as since they seemed to express positivity. Other words are negative or mixed sentiment, including *argh*, *ew*, *eek*, and *ugh*. The negative words were sometimes used figuratively in a positive or complementary context, such as in phrases like, "Arggghhh, you are so beautiful!" and "Uuugggh, size 12, I'm a 16 lol" .

Nine of the words were nouns, referring to things that were liked (Table 4). This was usually the influencer, referred to by name or as girl/gurl/bitch, but also included an influencer's dog (Toby). The term *bitch* was primarily directed from females to females in an apparently playful positive sentiment context. It can be a friendly insult or a term of affection, like gurl. The only general object that was commented on was hair, usually for a new hairstyle or colour. It is surprising that other common fashion and make-up objects were not in the list (e.g., lips, skin, dress, eyebrows).

Several words with many spelling variants had commenting conversational functions or started conversational moves (Table 5). Many of these words are extremely common, which makes them more likely to have spelling variants, other factors being equal. Two of the words (me, life) had

Table 3. Exclamations found in the 100 words with the most spelling variations.

Word	Comments	Spellings
yes	170808	337
ah	61801	328
yay	59385	293
omg	404200	292
wow	155107	262
yeah	52651	225
argh	2192	212
aw	65503	165
damn	34038	158
agh	2247	157
woah	7394	155
no	334387	144
god	109015	142
ugh	16915	142
ooh	13661	141
woohoo	3140	120
eek	3432	114
ew	5042	106
fuck	34726	106
aye	2515	104

Table 4. Nouns that are liked objects found in the 100 words with the most spelling variations.

Word	Comments	Spellings
girl	320339	352
gurl	14854	259
you	6597927	232
hair	431141	159
bitch	11906	100
Dina	33341	109
Nela	18866	119
Zoe	170546	104
Toby	1045	137

multiple different use contexts and did not fit a clear category. The conversational words show that repeated letters are not solely used to express approval or a positive opinion but can also be used to express what perhaps might be called cheerfulness (e.g., helllooo).

Finally, six words were present in the list due to single events: *coke* (milk and coke drink video), rain (makeup shade), *roar* (discussion of singer Katie Perry's song of that name), *squad* (influencer request for commenters to use this term), *help* (commenter worries about the physical safety or mental health of the influencer), *cage* (a popular large hamster cage introduced in a video).

Table 5. Other word functions found in the 100 words with the most spelling variations.

Word	Comments	Spellings	Function
same	258373	177	Conversational - agreement
first	259466	135	Conversational – first post
and	5414908	146	Conversational – emphasising second part of post
okay	47979	108	Conversational – starting or ending a sentence
hello	53201	192	Conversational – greeting
hi	246395	143	Conversational – greeting
hey	136179	130	Conversational – greeting
what	844110	218	Question or disbelief
please	499271	1397	Request
plz	43345	151	Request
me	1360534	106	Request (pick me); sentiment (slay me), reply
life	217318	121	Mix: slay my life, this gives me life, this video is life

DISCUSSION

This study is limited using a single corpus and a subjective judgement about what is a new sentiment slang term. The method might not work on other languages and especially those using pictograms or lacking orthographic pleonasm in their online culture. Although YouTube is international, it is banned in some countries, including China, and other video sharing sites are also popular elsewhere, so their comments might be used instead. The method may be less useful in the future if video sharing sites decrease in popularity in any countries. The results have not been assessed for comprehensiveness (recall) and it is possible that some new sentiment slang words have a more uniform spelling. The results also do not include some older slang sentiment words, such as wicked (13 variations) but this is to be expected in a contemporary corpus.

The results confirm that many, but not necessarily all, new sentiment slang words are frequently written with orthographic pleonasm. Moreover, this frequency is high enough compared to other words for it to be a useful filter to help identify words that are much more likely to be new sentiment terms than would a randomly chosen word. This provides evidence that the method works in practice, albeit with only one context tested so far: UK English, with a focus on the language of (probably) younger females.

The other types of words that are spelt in a wide variety of ways are also interesting from the perspective of sentiment analysis and the culture of online communication. For example, lexical algorithms to detect sentiment [26], [24] might consider incorporating rules to detect orthographic pleonasm in exclamations and booster words, where it seems particularly likely to affect the sentiment of the text.

CONCLUSIONS

The results show that the orthographic pleonasm method can identify new sentiment slang reasonably efficiently, given that 11% of the top 100 terms checked were valid. Additional terms were found with ranks 101-300 but it seems likely that sentiment slang would become much rarer lower in the list. This method should be considered as a supplement to existing methods, such as online slang or sentiment slang dictionaries [5], or as a replacement for languages or cultures lacking relevant dictionaries.

This paper also reports apparently the first evidence that it is possible to systematically extract slang from YouTube comments. Thus, other NLP projects focusing on informal text may also consider YouTube for this purpose.

The results have implications for the analysis of vocal spelling [19], [20], [22], showing that the orthographic pleonasm subtype is rich in comments on female UK influencers on YouTube and that spelling variants are not restricted to a few standard ones, but exhibit a huge range of variations. For example, the following variations of one word occur at least times: lysm:7592; lyssm:95; lysssm:81; lyssssm:77; lysmmmm:73; lysmmm:77; lysmm:52; lysssssm:43; lysssssm:38; lysmmmmm:30; llysm:26; lysssssssm:19; lysssssssssm:17; lysmmmmmm:14; lysmmmmmmm:10. The initialisms included in the results also extend previous analyses of vocal spelling because they conform to the repeated letters format but cannot be easily vocalised. These seem to be generalisations of the vocalisation technique: it is so familiar that the reader will understand that lysmmm emphasises lysm even though it does not make sense as a vocalisation.

As a practical step, organisations needing to identify new sentiment terms could repeat the exercise periodically (e.g., annually), recording the results of the manual checks each year so that repeat terms did not need to be checked again. This would produce a systematic method to ensure that the most popular sentiment-related terms were identified in a relatively efficient manner. The main time-consuming stages are curating the initial list of influencers and checking the top 100 terms for new sentiment slang, both of which will quicker after the first time because only changes need to be checked.

REFERENCES

1. D. S. Maylawati et al., “An improved of stemming algorithm for mining indonesian text with slang on social media,” *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 2018, pp. 1–6.
2. J. Coleman, *The life of slang*, OUP, 2012.
3. E. Cambria et al., “Sentiment analysis is a big suitcase,” *IEEE Intelligent Systems*, vol. 32, no. 6, 2017, pp. 74–80.
4. S. Correa and A. Martin, “Linguistic generalization of slang used in Mexican tweets, applied in aggressiveness detection,” *IberEval@ SEPLN*, 2018, pp. 119–127.
5. L. Wu, F. Morstatter, and H. Liu, “SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification,” *Language Resources and Evaluation*, vol. 52, no. 3, 2018, pp. 839–852.
6. N. Ul Haq et al., “USAD: An Intelligent System for Slang and Abusive Text Detection in PERSO-Arabic-Scripted Urdu,” *Complexity*, 2020.
7. T. Dalzell and T. Victor, *The concise new Partridge dictionary of slang and unconventional English*, Routledge, 2014.
8. A. Gupta et al., “SLANGZY: A fuzzy logic-based algorithm for English slang meaning Selection,” *Progress in Artificial Intelligence*, vol. 1, no. 8, 2019, pp. 111–121.

9. S. Wilson et al., “Urban dictionary embeddings for slang NLP applications,” *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4764–4773.
10. S. Dhuliawala, D. Kanojia, and P. Bhattacharyya, “Slangnet: A wordnet like resource for english slang,” *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 2016, pp. 4329–4332.
11. K. Matsumoto et al., “Slang feature extraction by analysing topic change on social media,” *CAAI Transactions on Intelligence Technology*, vol. 4, no. 1, 2019, pp. 64–71.
12. Z. Pei, Z. Sun, and Y. Xu, “Slang detection and identification,” *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 881–889.
13. F. M. Kundi et al., “Detection and scoring of internet slangs for sentiment analysis using SentiWordNet,” *Life Science Journal*, vol. 11, no. 9, 2014, pp. 66–72.
14. H. ElSahar and S. R. El-Beltagy, “A fully automated approach for Arabic slang lexicon extraction from microblogs,” *International conference on intelligent text processing and computational linguistics*, 2014, pp. 79–91.
15. N. Hossain, T. T. T. Tran, and H. Kautz, “Discovering political slang in readers’ comments,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018, pp. 612–615.
16. R. Komuda et al., “Recognizing and converting cockney rhyming slang for cyberbullying and crime detection,” *Proceedings of Language Sense on Computers IJCAI 2016 Workshop*, 2016.
17. F. Ren and K. Matsumoto, “Semi-automatic creation of youth slang corpus and its application to affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, 2015, pp. 176–189.
18. J. B. Walther and K. P. D’Addario, “The impacts of emoticons on message interpretation in computer-mediated communication,” *Social Science Computer Review*, vol. 19, no. 3, 2001, pp. 324–347.
19. J. Carey, “Paralanguage in computer mediated communication,” *18th Annual Meeting of the Association for Computational Linguistics*, 1980, pp. 67–69.
20. R. B. Harris and D. Paradice, “An investigation of the computer-mediated communication of emotions,” *Journal of Applied Sciences Research*, vol. 12, no. 3, 2007, pp. 2081–2090.
21. P. Shaw, “Spelling, accent and identity in computer-mediated communication,” *English Today*, vol. 24, no. 2, 2008, pp. 42–49.
22. M. A. Riordan and R. J. Kreuz, “Cues in computer-mediated communication: A corpus analysis,” *Computers in Human Behavior*, vol. 26, no. 6, 2010, pp. 1806–1817.
23. A. Hogenboom et al., “Exploiting emoticons in sentiment analysis,” *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 703–710.
24. M. Thelwall, K. Buckley, and G. Paltoglou, “Sentiment strength detection for the social web,” *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, 2012, pp. 163–173.
25. M.-C. Lien, P. A. Allen, and E. Ruthruff, “Multiple routes to word recognition: Evidence from event-related potentials,” *Psychological Research*, 2019, pp. 1–30.
26. M. Taboada et al., “Lexicon-based methods for sentiment analysis,” *Computational Linguistics*, vol. 37, no. 2, 2011, pp. 267–307.

Mike Thelwall is a professor of data science at the University of Wolverhampton, UK. His research interests include sentiment analysis and social media analysis. Thelwall received a PhD in pure mathematics from Lancaster University, UK. Contact him at m.thlewall@wlv.ac.uk.