Progress in Artificial Intelligence (2019) 8:111–121
https://doi.org/10.1007/s13748-018-0159-3

REGULAR PAPER

SLANGZY: a fuzzy logic-based algorithm for English slang meaning selection

Anshita Gupta[1] · Sanya Bathla Taneja[1] · Garima Malik[1] · Sonakshi Vij[1] · Devendra K. Tayal[1] · Amita Jain[2]

Received: 8 May 2018 / Accepted: 22 July 2018 / Published online: 7 August 2018

Abstract

The text present on online forums and social media platforms conventionally does not follow a standard sentence structure and uses words that are commonly termed as slang or Internet language. Online text mining involves a surfeit of slang words; however, there is a distinct lack of reliable resources available to find accurate meanings of these words. We aim to bridge this gap by introducing SLANGZY, a fuzzy logic-based algorithm for English slang meaning selection which uses a mathematical factor termed as "slang factor" to judge the accuracy of slang word definitions found in Urban Dictionary, the largest Slang Dictionary on the Internet. This slang factor is used to rank definitions of English slang words retrieved from over 4 million unique words on popular social media platforms such as Twitter, YouTube and Reddit. We investigate the usefulness of SLANGZY over Urban Dictionary to find meanings of slang words in social media text and achieve encouraging results due to recognizing the importance of multiple criteria in the calculation of slang factor in the algorithm over successive experiments. The performance of SLANGZY with optimum weights for each criterion is further assessed using the accuracy, error rate, F-Score as well as a difference factor for English slang word definitions. To further illustrate the results, a web portal is created to display the contents of the Slang Dictionary consisting of definitions ranked according to the calculated slang factors.

Keywords Slang · Fuzzy logic · Urban Dictionary · Text analysis

1 Introduction

Traditional natural language processing systems are designed to process formal language with standard grammatical structure. However, with the increasing popularity of social media websites, applications are being developed to gain insights from text mining on the Internet, widely in the area of sentiment analysis. Text present on online forums and social media platforms conventionally does not follow a standard sentence structure and uses words that are commonly termed as slang or Internet language. According to the Oxford Dictionary,[1] slang is "a type of language consisting of words and phrases that are regarded as very informal and are typically restricted to a particular context or group of people". The following examples illustrate commonly used slang words:

"Slammed": Having a large amount of work assigned to you that leaves you overwhelmed.
"Brexit": Britain's hypothetical exit from the European Union.
"Cool": Being fashionably attractive or impressive.

✉ Sonakshi Vij
sonakshi.vij92@gmail.com

Anshita Gupta
anshitag8@gmail.com

Sanya Bathla Taneja
sanyabt11@gmail.com

Garima Malik
annu.2353@gmail.com

Devendra K. Tayal
dev_tayal2001@yahoo.com

Amita Jain
amita_jain_17@yahoo.com

[1] Department of CSE, Indira Gandhi Delhi Technical University for Women, Delhi, India

[2] Department of CSE, Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India

[1] https://en.oxforddictionaries.com/.

Today, social media text and information retrieval is a major research area in natural language understanding. A significant proportion of text found on social media and other online sources constitutes informal language, and this cannot be ignored while developing new methods for natural language processing. Unfortunately, due to a dearth of reliable, up-to-date resources for such colloquialisms, data collected from social media websites are often termed as "noisy", and the preprocessing of text involves the rejection of words not included in resources such as WordNet [1]. This leads to discounting of words, which can be useful in most applications. For instance, in [2], short-noisy text like grt and g8t cannot be detected and handled properly. This holds true for most applications involving social media text wherein a large portion of relevant text is discounted simply due to a lack of definitive source for slang words and phrases.

The objective of this research is twofold. One, it presents a corpus of slang words currently used on popular social media websites such as Twitter,[2] YouTube[3] and Reddit,[4] along with information related to the word that can be used by both social media users and researchers. The definitions of each slang word are presented in order of relevance, judged by a slang factor. Categorization of slang words is also presented to maintain structure in the dataset. Web-based application of the Slang Dictionary represents the slang definitions in order of their slang factors. Two, it proposes a "slang factor", i.e. a mathematical factor calculated using fuzzy logic [3] on six criteria to judge the relevance and usefulness of each definition of slang word found on Urban Dictionary. This factor is useful in determining whether a slang word definition is relevant in the context of the word as well as validating the information found on the crowd-sourced dictionary. Fuzzification is done to make sure all possible criteria such as the number of upvotes, number of downvotes, average upvotes and more are considered while gauging the definition of the slang word. The details of the dataset creation and slang factor calculation are presented in Sect. 5.

We utilize Urban Dictionary[5] as a reference, which is currently the largest online dictionary of slang words with over seven million definitions at the end of 2014 [4]. As observed by [5], Urban Dictionary comprises nine properties, namely redundancy, self-reference, spelling/grammar, intra-colloquia, opinions, contrast, verbosity, multiple definitions and no parts of speech, that make it unusable as a lexical resource. Moreover, due to no internal or external validation, Urban Dictionary fails to control the quality of the definitions [6] and provides no information regarding the prevalence of the words. We aim to provide users and researchers with a reliable source for current English slang words by building upon and validating the information present in Urban Dictionary.

We investigate the effects of using SLANGZY in place of Urban Dictionary to find meanings of slang words in social media text and observe a significant improvement in the accuracy of definitions. The results presented in this paper highlight the usefulness of our resource over Urban Dictionary, both in the top definition provided and coverage of current slang words on social media platforms.

The rest of this paper is organized as follows. Section 2 surveys the related work. Section 3 presents the preliminaries. Section 4 describes the proposed algorithm. Section 5 elaborates our method used for corpus creation from Twitter, YouTube and Reddit as well as estimation of slang factor using fuzzy logic. Section 6 reports the experimental settings and results on definitions of slang words from both SLANGZY and Urban Dictionary, and, finally, Sect. 7 concludes the paper and discusses future directions and plans.

2 Related research

The processing and application of informal language have been of interest to many researchers. Dhuliawala et al. [6] present a WordNet-like resource for slang words and neologisms on the internet called SlangNet created using the top definitions from Urban Dictionary. Ni and Wang [7] use neural sequence learning techniques to build a model that generates explanations of English slang expressions given the context. The corpus is developed using the definitions and examples present in Urban Dictionary, and the model produces definitions of unseen words and phrases.

Research on English slang words has heavily focused on sentiment analysis of slang words. Wu et al. [8] put forth a sentiment dictionary of slang words for short-text sentiment classification using Urban Dictionary. The first version of this dictionary contains over 90,000 English slang words and phrases with sentiment scores. In fact, sentiment scoring of slang words is a popular topic of research. Kundi et al. [9] use SentiWordNet to detect and score Internet slangs found in Twitter data, although the detection and translation process ends with the top definition in Urban Dictionary.

Among the research on social media text, Chawla [10] performs sentiment analysis in Twitter where commonly used slangs, abbreviations and collocations are manually tagged. Often, the frequent spelling mistakes, abbreviations and slangs used in the tweets must be ignored as they are not included in the lexicons [2]. Pal and Saha [11] use a multi-step algorithmic and semi-supervised learning approach to detect jargon words in different e-texts. Other works on social media text involve analysis of the lexical variation in language found in online social media and contextual factors affecting the

[2] https://twitter.com/.

[3] https://www.youtube.com/.

[4] https://www.reddit.com/.

[5] https://www.urbandictionary.com/.

number of out-of-vocabulary words in micro-texts such as Twitter messages. Out-of-vocabulary terms that are not part of the common written English, lexicon comprises a significant portion of the dataset evaluated [12].

Comparison of noisy text from different social media sources based on language, lexical composition and grammaticality shows that social media text is less grammatical than edited text with high rates of out-of-vocabulary words [13]. Text normalization of noisy social media text such as by [14, 15] aims to identify and correct the lexical transformations that led to the formation of OOV words. Our approach considers these OOV words to be a part of Internet language if they are found in Urban Dictionary.

Urban Dictionary has long been the authority for people to keep up to date with the ever-changing Internet language, with over 30 million people viewing the site every month. However, assessing the viability of Urban Dictionary for use in conventional computational tasks through semantic similarity algorithms reveals certain issues which make it unsuitable for use with conventional dictionary-based methods [5].

Besides statistical and probabilistic approaches, fuzzy logic has been used for effective natural language processing in many applications. Adreevskaia and Bergler [16] present a dictionary-based Sentiment Tag Extraction Program to generate a fuzzy set of English sentiment-bearing words where fuzzy sentiment tags are assigned to all words in WordNet. Sun et al. [17] describe a fuzzy logic-based language processing model for speech recognition wherein fuzzy inference rules are applied for word prediction. Following the application of fuzzy set-based linguistic approach in decision analysis, the best solution among multiple definitions is selected based on various criteria with different linguistic weights [18]. To the best of our knowledge, this work is the first attempt at ranking the definitions of informal English words in Urban Dictionary using fuzzy logic.

3 Preliminaries

This section describes all preceding concepts and topics applied in this paper.

3.1 Urban Dictionary

Urban Dictionary is an online, crowd-sourced Slang Dictionary. It allows users to input definitions and examples of common slang words on the Internet and has a user upvote and downvote system, which decides the popularity of each definition. At the start of 2014, the dictionary featured more than seven million definitions of words, acronyms and phrases and 2000 definitions were added daily [4]. Each slang word in Urban Dictionary contains multiple definitions with each definition containing: upvotes, i.e. number of favourable votes given by users, downvotes, i.e. number of unfavourable votes given by users, example of usage and tags or related words.

3.2 Fuzzy logic

Fuzzy Logic is a methodology predicated on the idea that the "truthiness" of something can be expressed over a continuum [3, 19]. This implies that something is not true or false, rather partially true or partially false. Fuzzy logic variables have a truth value that ranges between 0 and 1 and is not constrained to the two truth values of classical propositional logic [20].

3.3 Fuzzy membership function

A fuzzy set (class) in X is characterized by a membership (characteristic) function $f(x)$ which associates with each point in X. A fuzzy set has pairs (U, m) where U is a set and m is a membership function which belongs in the interval [0, 1]. The reference set U is called universe of discourse, and for each $x \in U$, the value $m(x)$ is called the grade of membership x in (U, m) [3, 19].

The function $m = \mu(A)$ is called the *membership function* of the fuzzy set $A = (U, m)$.

For a finite set $U = \{x_1, \ldots, x_n\}$, the fuzzy set (U, m) is often denoted by $\left\{\frac{m(x1)}{x1}, \ldots, \frac{m(xn)}{xn}\right\}$.

The terms x taken together are the fuzzy set which can be used to describe the "fuzzy value" of a fuzzy variable. Each term has a membership function that defines how a value maps to the term on a scale of 0–1 [21]. 0 corresponds to the absence of membership in the category and 1 reflects the highest degree of membership. This paper uses the definitions of slang words as terms in the fuzzy set. The membership functions of various criteria are calculated and weights are assigned to each criterion. The summation of these gives the slang factor.

3.4 Types of fuzzy membership functions

Two types of membership functions of one dimension are used in this proposal, chosen per observed characteristics of the dataset:

3.4.1 Triangular membership function

The simplest membership functions are formed using straight lines. Of these, the simplest is the triangular membership

function [22] specified by three parameters $\{a, b, c\}$ as follows:

$$\text{triangle}(x; a, b, c) = \begin{cases} 0, x \leq a. \\ \frac{x-a}{b-a}, a \leq x \leq b. \\ \frac{c-x}{c-b}, b \leq x \leq c. \\ 0, c \leq x. \end{cases} \quad (1)$$

The triangular membership function gives the membership degree of each criterion depending on the parameters $\{a, b, c\}$(with $a < b < c$) which determine the coordinates of the three corners.

3.4.2 Sigmoidal membership function

The sigmoidal function [22], $sigmf(x, [a, c])$, as given in the following equation by $f(x, a, c)$ is a mapping on a vector x and depends on two parameters a and c. It is specified as:

$$f(x, a, c) = \frac{1}{1 + e^{-a(x-c)}} \quad (2)$$

The sign of the parameter a determines if the sigmoidal membership function is inherently open to the right or to the left, and thus is appropriate for representing the "very large" values in our dataset. The sigmoid function gives the membership degree of each criteria depending on the values of a and c.

4 Proposed algorithm

The algorithm SLANGZY is used to calculate the slang factors for all definitions of different slang words present in the dataset (Table 1).

The algorithm applies the fuzzy membership functions from Eqs. (1) and (2) in various functions depending on the criteria.

The definitions with maximum slang factors calculated for slangs using the SLANGZY algorithm are the top definitions according to our proposal.

5 Implementation

To implement the SLANGZY algorithm, the following course of action is followed. It includes three steps, namely data collection and preprocessing, slang database creation followed by fuzzification and slang factor generation.

1. Data collection and preprocessing
   Sentences are collected from three popular textual social media websites Twitter,[6] YouTube[7] and Reddit[8] (Fig. 1). These were chosen due to the substantial research in social media text done using these sources as well as the high proportion of slang usage on these websites. The first step is to clean the data and remove particular problems associated with text found on the specified social media websites such as usernames, escaping HTML characters, stop words, punctuations, URLs, ellipses, emoticons and other special characters [23]. Total sentences post the cleaning steps are 3.36 million. After tokenization, approximately 4 million unique words occurring more than once are taken to avoid chances of spelling errors.
2. Slang database creation
   The slang database contains slangs that have been divided into three cases using PyEnchant,[9] WordNet and Urban Dictionary, namely
   Case 1:
   These are words present in the English Dictionary, WordNet as well as Urban Dictionary. These may be slangs with multiple senses or non-slang words. For example, the word "butterfly".
   Definition in Wordnet: (noun), "diurnal insect typically having a slender body with knobbed antennae and broad colorful wings"
   Definition in Urban Dictionary: "the feeling you get in your stomach when the person you have a crush on walks into room".
   Case 2:
   These are words present in English Dictionary and Urban Dictionary, but not in WordNet. These are slang words with multiple senses or slang words that have been recognized in formal dictionaries. For example, the word "hangup".
   Case 3:
   These are words only present in Urban Dictionary. These are confirmed to be slang words in the database. For example, the word "Brexit".
   Definition in Urban Dictionary: "Britain's hypothetical exit from the European Union".
   For each slang word in the dataset, all the above information is extracted using the Urban Dictionary API.[10] The database now contains the following information for each word: slang word, definition with example (may be mul-

[6] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview.

[7] https://developers.google.com/youtube/v3/.

[8] https://www.reddit.com/r/all/comments/.json?.

[9] www.pythonhosted.org/pyenchant/.

[10] api.urbandictionary.com/v0/define?.

Table 1 Proposed SLANGZY algorithm

Algorithm Input: $\{key: value\}$ pairs with $key = slang\ word$

Algorithm Output: Slang Factor derived by Fuzzy Membership Functions

Method:

1. Slang words with their definitions will be input. Each definition has upvotes, downvotes and ratio values attached to it.
2. Define the universe discourse, i.e., definitions and fuzzy sets.
3. Find the maximum values of upvotes and number of words in the dataset.
4. Define fuzzy membership functions for all criteria as:
   - 4.1. ***U**: Sigmoid Minimum Upvotes Membership Function:*
     Starting minimum value = 100
     Increment value = 1
     Ending maximum value = Maximum Upvotes for all definitions of key
     Sigmoid function values: $c = \frac{(Maximum\ Upvotes)}{2}, a = \frac{1}{c}$
   - 4.2. ***R**: Sigmoid Ratio Membership Function*
     Starting minimum value = 0
     Increment value = 1
     Ending maximum value = Maximum Ratio for all definitions of key
     Sigmoid function values: $c = \frac{(Maximum\ Ratio)}{2}, a = \frac{1}{c}$
   - 4.3. ***D**: Sigmoid Difference Membership Function*
     Starting minimum value = 1
     Increment value = 1
     Ending maximum value = Upvotes of current definition
     Sigmoid function values: $c = \frac{(Current\ Upvotes)}{2}, a = \frac{1}{c}$
   - 4.4. ***W**: Triangular Number of Words Membership Function*
     Starting minimum value = 1
     Ending maximum value = Maximum Number of Words for all de finitions of key
     Triangular function values: $5, 10, 100$
   - 4.5. ***A**: Boolean Average Upvotes Membership Function*
     If greater than average: 1
     Else: 0
5. Calculate each fuzzy membership function (U, R, D, W, A) for an item in the fuzzy set.
6. Assign weights to each membership function as w_i:
   $U \rightarrow 0.46$
   $R \rightarrow 0.6$
   $D \rightarrow 0.4$
   $W \rightarrow 0.14$
   $A \rightarrow 0.4$
   Maximum value of Slang Factor is 2.
7. $Slang\ Factor = \sum_{i=1}^{6}(w_i * c_i)$ (3)
   where $c_i = \{U, R, D, W, A\}$ and each $c_i \in [0,1]$.
8. Select the fuzzy member with the maximum value of Slang Factor as the best definition for a given slang word.
9. Repeat steps 5 to 8 for all available fuzzy set items in the input dataset.
10. Stop.

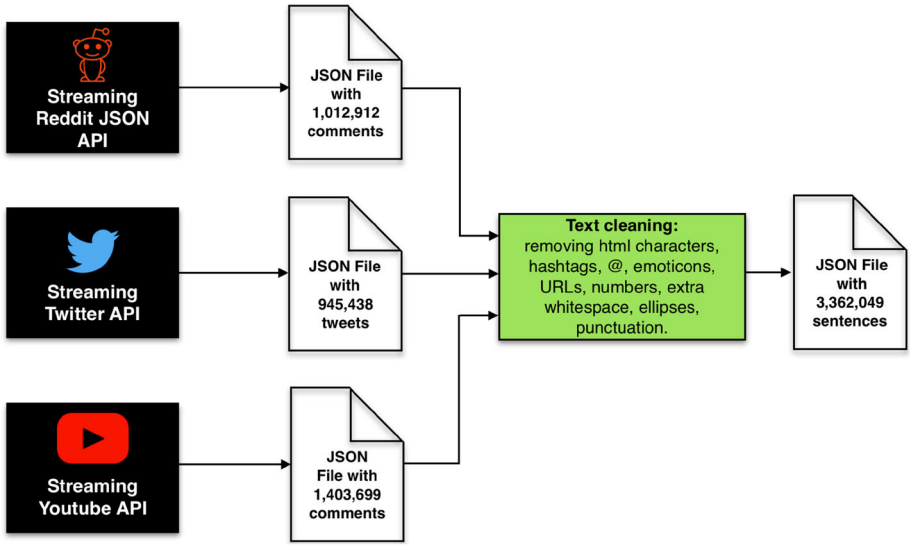

Fig. 1 Data collection and preprocessing

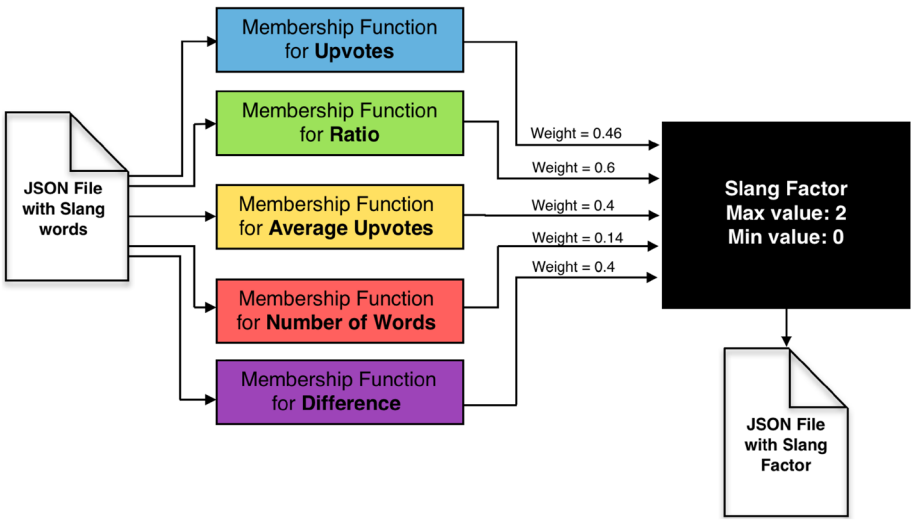

Fig. 2 Slang factor generation

tiple), upvotes, downvotes, ratio of upvotes to downvotes and tags.

3. Fuzzification and slang factor generation
   Urban Dictionary uses the difference between the number of upvotes and downvotes to decide the top definition for each word in the dictionary (Fig. 2). As observed, this alone does not guarantee the top definition to be useful or even accurate. We use the following five criteria to judge the usefulness and relevance of the slang words in the database:

   1. Number of upvotes
      The number of upvotes of each definition. The higher the upvotes, more is the popularity and usage of the slang word.
   2. Ratio of upvotes to downvotes
      This is an important factor, which indicates the popularity of the slang word. Favourable values are greater than 2.
   3. Number of words in definition
      The number of words in each definition should not be too low or too high. If too low, the meaning may be of no use or false. If it is too high, there might be irrelevant information added.
   4. Average upvotes
      Average upvotes criterion is chosen to make sure that the upvotes for each definition are greater than the average upvotes of all definitions for a given slang word.
   5. Difference between upvotes and downvotes
      Urban Dictionary defines the top definition of a slang word as the definition with the maximum difference

Table 2 Membership functions and final weights of criteria used for slang factor calculation

Criteria	Membership function	Weight (W_i)
Number of upvotes	Sigmoidal	0.46
Ratio of upvotes to downvotes	Sigmoidal	0.6
Number of words in definition	Triangular	0.14
Average upvotes	Binary (0/1)	0.4
Difference b/w upvotes and downvotes	Sigmoidal	0.4

between upvotes and downvotes. This is also taken as one of the criteria.

5.1 Slang factor

We propose a factor known as "slang factor" to indicate the usefulness of each definition of slang word. This factor is calculated using fuzzy membership functions from the Python library scikit-fuzzy.[11] This gives a value between 0 and 1 for each variable and depending on the importance, weights were assigned experimentally to each criterion. The cumulative of the membership values and the weights is calculated to give the slang factor:

$$\text{Slang Factor} = \sum_{i=1}^{6} (w_i * c_i) \tag{3}$$

where w_i is the weight (Table 2) assigned to each criterion and c_i is the membership value.

6 Evaluation and results

We evaluate the performance of SLANGZY with various sample test cases to judge the importance of each criterion. Successive experiments indicate that the ratio of upvotes to downvotes is more important in selecting an appropriate slang definition than other criteria such as the number of upvotes, the difference between upvotes and downvotes and number of words. Thus, the quality of top definition of slang word can be significantly improved by considering each criterion's importance while calculating the slang factor and ranking the definitions of each slang word with SLANGZY.

The weights assigned to the criteria are adjusted experimentally to assess the degree of importance of each criterion while ranking the slang word definitions. The results of the experiments are displayed in Table 3 with the weights considered, accuracy and error rate of each sample. We observe significant improvement over the multiple samples evaluated, finally giving the optimum weights for calculation of slang factor as shown in sample IV.

Table 3 Experimental weights of criteria used for slang factor in four samples

Sample	Weights (*U, R, D, W, A*)	Accuracy (%)	Error rate (%)
I	0.8, 0.6, 0.5, 0.2, 0.5	74.49	25.50
II	0.6, 0.7, 0.6, 0.2, 0.5	86.82	13.18
III	0.6, 0.8, 0.5, 0.2, 0.5	85.65	14.35
IV	0.46, 0.6, 0.4, 0.14, 0.4	88.44	11.55

After finding the top definition according to Urban Dictionary and SLANGZY, the slang words are divided into separate categories:

Category 1: Same correct top definition according to Urban Dictionary and SLANGZY.
Category 2: SLANGZY gives correct top definition.
Category 3: Urban Dictionary gives correct top definition.
Category 4: Both SLANGZY and Urban Dictionary give incorrect top definition.

Confusion matrix (CF) [24], also called error matrix, is used to present the system performance. CF presents the actual and predicted results in the form of TP, TN, FP and FN, which stands for true positive, true negative, false positive and false negative, respectively.

In the following equations [25], TP (True Positives) and FN (False Negatives) are the number of known correct top definitions predicted correctly by SLANGZY and incorrectly by SLANGZY, respectively. Correspondingly, TN (True Negatives) and FP (False Positives) are the number of known incorrect top definitions predicted to be incorrect by SLANGZY and correct by SLANGZY, respectively [26].

$$\text{Accuracy } (A) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{4}$$

$$\text{Error Rate}(E) = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{5}$$

$$\text{Specificity}(S) = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{6}$$

$$F - \text{Score}(F) = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{7}$$

The performance of SLANGZY with optimum weights is given in Table 4, calculated using Eqs. (4)–(7) for two

[11] https://github.com/scikit-fuzzy/scikit-fuzzy.

Table 4 Results calculated for optimum weights

Performance measures (%)	Case 1	Case 2
Accuracy (A)	88.44	72.35
Error rate (E)	11.55	27.64
Specificity (S)	73.43	73.43
F-Score	92.61	71.18

Table 5 Cases of slang words where Urban Dictionary fails to give useful top definitions and their slang factors

Slang	Top definition slang factor	Average slang factor	Difference factor
emo	1.170	0.672	0.498
Killing	1.226	0.522	0.704
gf	1.215	0.639	0.576
Selfie	1.135	0.549	0.585
Hashtag	1.115	0.594	0.521
Vis-à-vis	1.177	0.364	0.813
Horseradish	1.172	0.548	0.623
Admin	1.181	0.674	0.507
Brexit	1.238	0.648	0.589

cases over the test data. Case 1 includes the categories of slang words where both SLANGZY and Urban Dictionary give the same correct top definition. In Case 2, these slang words are excluded from the evaluation. The high accuracies and low error rates indicate that several factors in addition to the difference between upvotes and downvotes contribute to the selection of the most accurate definition of a slang word. Choosing the best definition by taking the contribution of each criterion with its importance is crucial in deciding the best meaning and getting correct definitions with an accuracy of 88.44 percentage.

Using fuzzy membership functions in SLANGZY to evaluate the slang factor takes into account the contribution of multiple criteria while calculating the slang factor for each slang word definition. Thus, the fuzzy logic-based algorithm allows for the inclusion of human assessments in computing problems and provides an effective means for conflict resolution of multiple criteria and better assessment of options [27]. A few examples of slang words along with the slang factors of the top definition of each word are presented in Table 5. We also measure the difference factor of each slang word as the difference between the slang factor value of the best definition by SLANGZY and the average slang factor value of the remaining definitions. This constitutes the "degree of preferability" of the chosen definition, i.e. the amount by which the chosen solution is preferable to the other [28].

Table 6 shows examples of English slang words where the definition with the highest slang factor is more accurate and intuitively more understandable than the top definition of the same word in Urban Dictionary.

6.1 Portal for Slang Dictionary

A web portal (Fig. 3) is created to explore the contents of the dictionary and display the results of the database and slang factors.

Users can enter English slang words to search the dictionary which presents the following information or each word:

- Definition
- Upvotes
- Downvotes
- Ratio of upvotes to downvotes
- Difference between upvotes and downvotes
- Slang factor
- Link to access the word on Urban Dictionary

The portal offers an interactive way to access the slang database where users can judge the relevance and usefulness of each definition of the slang word found in Urban Dictionary. The definitions are presented in decreasing order of slang factor such that the definition with maximum slang factor is the "top definition".

7 Conclusion and future work

This paper presents a fuzzy logic-based approach to select the most accurate definition of English slang words present in Urban Dictionary. The algorithm, SLANGZY, succeeds in normalizing the unstructured meanings of slang words in Urban Dictionary to provide a resource which ranks the slang definitions and extracts the most relevant meaning. This is done using the proposed "slang factor" based on five relevant criteria for definitions of social media jargon. The comprehensive dataset created achieves an accuracy of 88.44% while selecting the best definition in comparison with the top definition of the same word in Urban Dictionary. To supplement the conventional NLP applications which primarily work with the meaning of textual data extracted from the internet, SLANGZY can provide accurate definitions of non-standard English words to not hinder information retrieval through social media text. To extenuate the effects of slang in the processing of natural language text using conventional approaches, users can access the web portal and get the most accurate top definition of the slang words. Currently, our corpus has approximately one million English slang words along with their slang factors. We aim to use SLANGZY in the future to define meaningful relationships among the

Table 6 Examples of slang words where Urban Dictionary fails to give useful top definitions

Sentence	Slang used	UD top definition	SLANGZY top definition	Correct definition
These are the people that force themselves to be emo, try to dramatize their life	emo	A musical genre/scene that has almost 1000 definitions in Urban Dictionary most of which are making fun of it	Genre of softcore punk music that integrates unenthusiastic melodramatic 17 year olds	Genre of softcore punk music that integrates unenthusiastic melodramatic 17 year olds
I saw their band last night, they were killing it	Killing	Chuck Norris' alternative to going hunting	Doing exceptionally well at something. Usually followed by "it"	Doing exceptionally well at something. Usually followed by "it"
Your gf has good sense of humour.	gf	George Foreman (Grill)	1. Chatting: Short for "Girl Friend""2. Gaming: Jedi-Outcast: "Good Fight" usually said after a Lightsaber duel	1. Chatting: Short for "Girl Friend" 2. Gaming: Jedi-Outcast: "Good Fight" usually said after a Lightsaber duel
Mounting the camera on a tripod, or resting it securely on a table, and using a timer to take the photograph is not a selfie	Selfie	The act of performing fellatio on oneself	A picture taken of yourself that is planned to be uploaded to Facebook, Myspace or any other sort of social networking website. A selfie is usually accompanied by a kissy face or the individual looking in a direction that is not towards the camera	A picture taken of yourself that is planned to be uploaded to Facebook, Myspace or any other sort of social networking website. A selfie is usually accompanied by a kissy face or the individual looking in a direction that is not towards the camera
The hashtag Arctic Monkeys is trending on Twitter	Hashtag	A ridiculous replacement for the "pound sign"	If you have been on Twitter, you may have seen a "hashtag". To put it simply, a hashtag is simply a way for people to search for tweets that have a common topic and to begin a conversation	If you have been on Twitter, you may have seen a "hashtag". To put it simply, a hashtag is simply a way for people to search for tweets that have a common topic and to begin a conversation
They were now vis-à-vis the most famous painting in the Louvre	Vis-à-vis	A coconut	In relation to or in comparison with	In relation to or in comparison with
That referee call was horseradish, if you ask me	Horseradish	Like mayonnaise, it's not an instrument	(adj.) Utterly inaccurate used to express an opinion of disbelief	(adj.) Utterly inaccurate used to express an opinion of disbelief
Can I just get a straight answer from admin to whether or not we can write on the forum?	Admin	A satanic race of pure evil, who rule the Internet through a Monarchistic feudelic system	An administrator of a computer system, an online forum etc.	An administrator of a computer system, an online forum etc.
After the brexit vote, British prime minister, David Cameron announced his resignation	brexit	British slang for coitus interruptus.	A word describing Britain's hypothetical exit from the European Union. A [portmanteau] of Britain and exit	A word describing Britain's hypothetical exit from the European Union. A [portmanteau] of Britain and exit

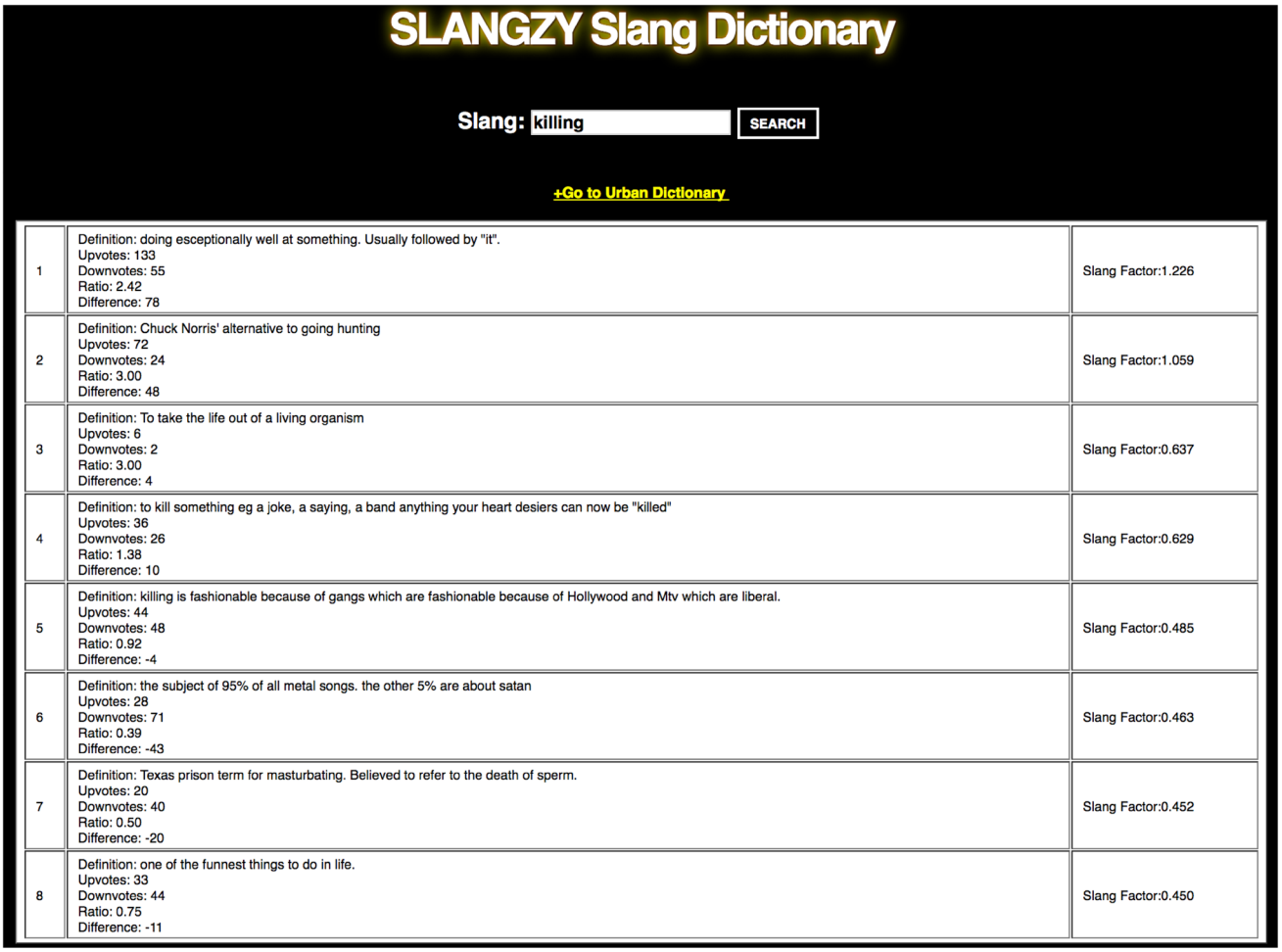

Fig. 3 SLANGZY Slang Dictionary portal

slang words and find associations derived from contextual information. The coaction of this resource along with algorithms used for query expansion, word sense disambiguation and sentiment analysis will make a powerful system for efficient natural language processing and understanding.

References

1. Fellbaum, C.: WordNet. Wiley, New York (1998)
2. Mukherjee, S., Bhattacharyya, P.: Sentiment analysis in twitter with lightweight discourse analysis. Proc. COLING **2012**, 1847–1864 (2012)
3. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
4. Wortham, J.: A lexicon of instant argot. The New York Times, 3 Jan 2014
5. Swerdfeger, B.A.: Assessing the viability of the Urban Dictionary as a resource for slang. http://www.cs.ubc.ca/ (2012)
6. Dhuliawala, S., Kanojia, D., Bhattacharyya, P.: SlangNet: a WordNet like resource for English Slang. In: LREC (2016)
7. Ni, K., Wang, W.Y.: Learning to explain non-standard English words and phrases. arXiv preprint arXiv:1709.09254 (2017)
8. Wu, L., Morstatter, F., Liu, H.: SlangSD: building and using a sentiment dictionary of slang words for short-text sentiment classification. arXiv preprint arXiv:1608.05129 (2016)
9. Kundi, F.M., Ahmad, S., Khan, A., Asghar, M.Z.: Detection and scoring of internet slangs for sentiment analysis using SentiWordNet. Life Sci. J. **11**(9), 66–72 (2014)
10. Chawla, K., Ramteke, A., Bhattacharyya, P.: Iitb-sentiment-analysts: participation in sentiment analysis in twitter semeval 2013 task. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol. 2, pp. 495–500 (2013)
11. Pal, A.R., Saha, D.: Detection of slang words in e-data using semi-supervised learning. arXiv preprint arXiv:1702.04241 (2015)
12. Gouws, S., Metzler, D., Cai, C., Hovy, E.: Contextual bearing on linguistic variation in social media. In: Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, pp. 20–29 (2011)
13. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L.: How noisy social media text, How different social media sources? In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 356–364 (2013)
14. Cook, P., Stevenson, S.: An unsupervised model for text message normalization. In: Proceedings of the Workshop on Computational

Approaches to Linguistic Creativity, Association for Computational Linguistics, pp. 71–78 (2009)
15. Han, B., Baldwin, T.: Lexical normalisation of short text messages: Makn sens a# twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, Association for Computational Linguistics, pp. 368–378 (2011)
16. Adreevskaia, A., Bergler, S.: Mining wordnet for a fuzzy sentiment: sentiment tag extraction from wordnet glosses. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
17. Sun, J., Karray, F., Basir, O., Kamel, M.: Fuzzy logic-based natural language processing and its application to speech recognition. In: 3rd WSES International Conference on Fuzzy Sets and Systems (2002)
18. Tong, R.M., Bonissone, P.P.: A linguistic approach to decision making with fuzzy sets. In: Readings in Fuzzy Sets for Intelligent Systems, pp. 777–784 1993
19. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, New Jersey (1996)
20. Mandal, S.N., Choudhury, J.P., Chaudhuri, S.B.: In search of suitable fuzzy membership function in prediction of time series data. Int. J. Comput. Sci. Issues **9**(3), 293–302 (2012)
21. Zadeh, L.A.: Calculus of fuzzy restrictions. In: Zadeh, L.A., Fu, K.-S., Tanaka, K., Shimura, M. (eds.) Fuzzy Sets and their Applications to Cognitive and Decision Processes, pp. 1–40. Academic Press Inc., New-York (1975)
22. Zhao, J., Bose, B.K.: Evaluation of membership functions for fuzzy logic controlled induction motor drive. In: Proceedings of 28th Annual IECON Conference, vol. 1, IEEE Industrial Electronics Society, pp. 229–234 (2002)
23. Provost, F.J., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. ICML **1998**(98), 445–453 (1998)
24. Mukherjee, S., Malu, A., et al.: TwiSent: a multistage system for analyzing sentiment in twitter. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, Hawaii, USA, ACM, pp. 2531–2534 (2012)
25. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**(4), 427–437 (2009)
26. Su, B.H., Shen, M.Y., Esposito, E.X., Hopfinger, A.J., Tseng, Y.J.: In silico binary classification QSAR models based on 4D-fingerprints and MOE descriptors for prediction of hERG blockage. J. Chem. Inf. Model. **50**(7), 1304–1318 (2010)
27. Singh, H., Gupta, M.M., Meitzler, T., Hou, Z.G., Garg, K.K., Solo, A.M.G., Zadeh, L.A.: Real-life applications of fuzzy logic. Adv. Fuzzy Syst. 2013, Article ID 581879, 3p (2013)
28. Bonissone, P.P.: A fuzzy sets based linguistic approach: theory and applications. In: Proceedings of the 12th Conference on Winter Simulation, IEEE Press, pp. 99–111 (1980)