Urban Dictionary Sentiment Analysis

Introduction to NLP
AIT 526-004 Group 1
George Mason University
September 22nd, 2023

Introduction

Understanding the attitudes and opinions behind online is a crucial skill for many people. From creating policy to creating products or analyzing customer feedback, there are many situations in which quickly and accurately determining public sentiment is crucial in making quick and effective decisions. The most expansive resource of opinions is on social media sites such as Twitter and Facebook, where one can find thoughts on nearly any topic. However, due to the informal nature of social media, many users use slang words. Because these are colloquial words, there are limited resources that can account for their impact on sentiment.

Our project hopes to provide a resource that catalogs the sentiment values of slang words. This tool would help provide more precise measures of sentiment across social media. This will be done by utilizing Urban Dictionary, a sizeable crowd-sourced repository of slang words and their definitions. By creating a source to reference the sentiment behind slang words that are not encountered often otherwise, we can have a better way to interpret the sentiment behind text found on social media. We can supply a helpful resource that can add to the existing toolkits for sentiment analysis by documenting the feeling of these words.

Related Work

There have been numerous forays into sentiment analysis of online texts that display the benefits of such endeavors and the benefits of understanding the attitude behind slang words.

One such example is the study done by Qi and Shabrina in 2023. They used Twitter data to analyze the general response to the third Covid-19 lockdown in England. Since Twitter is a natural place for discourse on social media, the corpora examined in this study provided good insight into how users felt during this time. Information like this can help determine public policy by informing policymakers of the public sentiment. It showcases the benefits of sentiment analysis of online discourse. Thus, a better understanding of online jargon can only serve to aid these efforts.

Singh and Kumari (2016) describe an issue with sentiment analysis of tweets. They regularly come across slang words that do not have ascribed sentiment, thus making understanding the overall emotion behind the tweet difficult when performing sentiment analysis. They suggest a model in which they perform extensive pre-processing to prepare the tweet for analysis. Among the pre-processing tools is a step that requires determining the importance of a newly encountered slang word, and if that word is determined to have significance on the meaning of the tweet, they further use context to estimate the sentiment of the word. As Singh and Kumari described, slang words come up regularly when analyzing social media posts, and this study highlights the necessity of having a readily available tool to provide the sentiment of slang words.

Wu, Morstatter, and Liu (2016) similarly sought to create a dictionary that analyzes slang words using Urban Dictionary as their source of words. They used Twitter to examine these words' usage and derive the sentiment associated with them by gleaning the feelings behind the tweets. This is a practical approach generally, but when the definition of a word is readily available, it is more straightforward to determine the general sentiment of the word using the provided definition rather than using the context in tweets.

These previous studies highlight the need for a resource like the one this program seeks to create. Qi and Shabrina show the practical impact of sentiment analysis on online posts. Singh and Kumari describe how a lack of insight into the sentiment of slang words can depreciate the effectiveness of sentiment analysis. Wu, Morstatter, and Liu also saw the benefits of creating a dictionary of slang sentiment, but their approach focused on slang words in the context of tweets. Our project aims to combine the rationale of all these previous studies with the added benefit of access to Urban Dictionary definitions to create an accurate interpretation of these words and their sentiment.

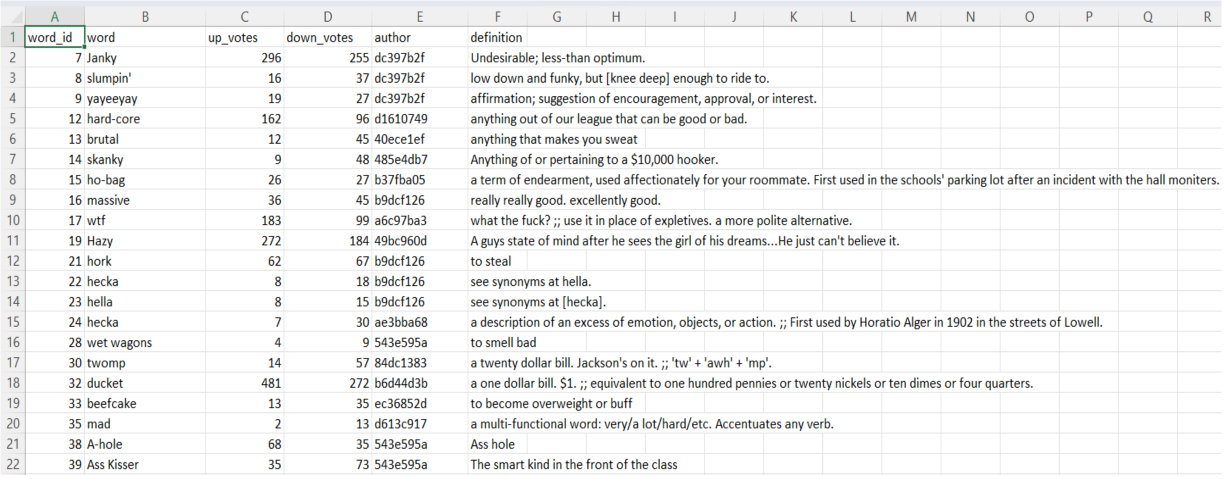
Objective

The main objective of our project is to utilize Natural Language Processing topics learned in class, to create an innovative model cable of categorizing words and phrases from Urban Dictionary as either positive or negative by leveraging their definitions and usage examples. There will be a multi-step process to achieve our goal, the first being collection where we will use an existing Kaggle dataset with numerous Urban Dictionary entries. We will use text processing techniques to clean and structure the data to ensure our model receives quality input. The focus of our project is to apply sentiment analysis algorithms to a data set of unique words. These algorithms will enable our model to not only identify terms as positive or negative but also understand the depth of the sentiment within these expressions.

Ultimately, our project aims to construct a versatile and insightful NLP resource that not only aids in comprehension of sentiment within the urban language, but also acts as a valuable asset for NLP research. Our NLP model can contribute to a broader field of language understanding, providing researchers with a tool to use when exploring our constantly evolving dictionaries. This can go beyond social media to act as an addition to any lexicon-based sentiment analysis approach, as it will be beneficial in any scenario involving text where slang may be encountered.

Selected Dataset

The “Urban Dictionary Words And Definitions” data set from Kaggle will be used in this project (Kulkarni, 2016). It consists of over 2.5 million entries from Urban Dictionary and six attributes. e “word_id” column is for usage in the Urban Dictionary API. The “word” column is the entry from Urban Dictionary that will be classified as positive or negative. The “up_votes” and “down_votes” columns are numerical attributes that display the number of thumbs up or down that the phrase and definition received on the website. The “author” column displays a hash of the user who uploaded the definition. Lastly, the “definition” column gives us the text we will analyze to provide a sentiment score for the slang word.



word_id	word	up_votes	down_votes	author	definition
7	Janky	296	255	dc397b2f	Undesirable; less-than optimum.
8	slumpin'	16	37	dc397b2f	low down and funky, but [knee deep] enough to ride to.
9	yayeeyay	19	27	dc397b2f	affirmation; suggestion of encouragement, approval, or interest.
12	hard-core	162	96	d1610749	anything out of our league that can be good or bad.
13	brutal	12	45	40ece1ef	anything that makes you sweat
14	skanky	9	48	485e4db7	Anything of or pertaining to a $10,000 hooker.
15	ho-bag	26	27	b37fba05	a term of endearment, used affectionately for your roommate. First used in the schools' parking lot after an incident with the hall moniters.
16	massive	36	45	b9dcf126	really really good. excellently good.
17	wtf	183	99	a6c97ba3	what the fuck? ;; use it in place of expletives. a more polite alternative.
19	Hazy	272	184	49bc960d	A guys state of mind after he sees the girl of his dreams...He just can't believe it.
21	hork	62	67	b9dcf126	to steal
22	hecka	8	18	b9dcf126	see synonyms at hella.
23	hella	8	15	b9dcf126	see synonyms at [hecka].
24	hecka	7	30	ae3bba68	a description of an excess of emotion, objects, or action. ;; First used by Horatio Alger in 1902 in the streets of Lowell.
28	wet wagons	4	9	543e595a	to smell bad
30	twomp	14	57	84dc1383	a twenty dollar bill. Jackson's on it. ;; 'tw' + 'awh' + 'mp'.
32	ducket	481	272	b6d44d3b	a one dollar bill. $1. ;; equivalent to one hundred pennies or twenty nickels or ten dimes or four quarters.
33	beefcake	13	35	ec36852d	to become overweight or buff
35	mad	2	13	d613c917	a multi-functional word: very/a lot/hard/etc. Accentuates any verb.
38	A-hole	68	35	543e595a	Ass hole
39	Ass Kisser	35	73	543e595a	The smart kind in the front of the class

Proposed System

This program will take in the words and perform sentiment analysis on the associated definition to create a dictionary that documents the sentiment of the slang words on Urban Dictionary.

First, the program will input the words and definitions from the provided data set. Then, it must pre-process the definition text to clean up the words. This will consist of tokenizing the definitions, ensuring the text is in all lowercase, removing punctuation and stop words, stemming them, and then tagging them with their parts of speech.

Once the definitions have been formatted to be analyzed, we will perform sentiment analysis on them. This will consist of using available tools to determine the sentiment of the words in the definition, and then derive the overall sentiment of the slang word.

This will be done for all the words and definitions provided by Urban Dictionary. Given that Urban Dictionary serves as an expansive catalog of slang words and their meanings, this will result in a helpful repository for analyzing online text sentiment.

Proposed Development Platform

This project will leverage Python and its associated libraries. Python's ease of use and existing data science tools make it a strong choice for this project's natural language processing tasks. Its helpful libraries include Natural Language Toolkit (NLTK) and pandas.

Pandas is an open-source Python library for data manipulation and analysis. It provides easy-to-use data structures and functions for working with structured data, making it the best option for managing the data in the data set.

NLTK is an extensive tool that can perform the text processing needs of this program. It has the capacity to tokenize the definitions, remove punctuation and stop words, tag the parts of speech, and lemmatize words. It can also aid in sentiment analysis of the definitions.

References

Kulkarni, R. Urban Dictionary Words And Definitions. (2019). https://www.kaggle.com/datasets/therohk/urban-dictionary-words-dataset

Singh, T., Kumari, M. Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science. 89. (2016). https://doi.org/10.1016/j.procs.2016.06.095

Qi, Y., Shabrina, Z. Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. *Soc. Netw. Anal. Min.* 13, 31 (2023). https://doi.org/10.1007/s13278-023-01030-x

Wu, L., Morstatter, F., Liu, H. SlangSD: Building and Using a Sentiment Dictionary of Slang Words for Short-Text Sentiment Classification. (2016). https://doi.org/10.48550/arXiv.1608.05129