Leveraging time-dependent lexical features for offensive language detection

Barbara McGillivray
King's College London &
The Alan Turing Institute
barbara.mcgillivray@kcl.ac.uk

Malithi Alahapperuma
Independent Researcher
malithi.alahapperuma@gmail.com

Jonathan Cook
University of Oxford
jonathan.cook2@eng.ox.ac.uk

Chiara Di Bonaventura
King's College London
chiara.di_bonaventura@kcl.ac.uk

Albert Meroño-Peñuela
King's College London
albert.merono@kcl.ac.uk

Gareth Tyson
Hong Kong University of
Science and Technology (GZ)
gtyson@ust.hk

Steven R. Wilson
Oakland University
stevenwilson@oakland.edu

Abstract

We present a study on the integration of time-sensitive information in lexicon-based offensive language detection systems. Our focus is on Offenseval sub-task A, aimed at detecting offensive tweets. We apply a semantic change detection algorithm over a short time span of two years to detect words whose semantics has changed and we focus particularly on those words that acquired or lost an offensive meaning between 2019 and 2020. Using the output of this semantic change detection approach, we train an Support Vector Machine (SVM) classifier on the Offenseval 2019 training set. We build on the already competitive SINAI system submitted to Offenseval 2019 by adding new lexical features, including those that capture the change in usage of words and their association with emerging offensive usages. We discuss the challenges, opportunities and limitations of integrating semantic change detection in offensive language detection models. Our work draws attention to an often neglected aspect of offensive language, namely that the meanings of words are constantly evolving and that NLP systems that account for this change can achieve good performance even when not trained on the most recent training data.

1 Introduction

The task of automatic detection of offensive language has attracted considerable attention in the Natural Language Processing (NLP) community recently. Policy makers and online platforms can leverage computational methods of offensive language detection to oppose online abuse and online harm at scale. These methods can also support computational social science and linguistics research in identifying innovative ways in which individuals and groups express offense online (Garland et al., 2020). The last two editions of the OffensEval shared task, organised as part of the SemEval competition, have offered a platform for assessing the state of the art in this area. Existing methods for automatic offensive language detection have been tested on a different large-scale annotated datasets, most notably the Offensive Language Identification Dataset (OLID) used in OffensEval 2019 (Zampieri et al., 2019a) and the Semi-Supervised Offensive Language Identification Dataset (SOLID) from OffensEval 2020 (Rosenthal et al., 2020).

The most effective methods proposed so far typically rely on ensembles of very large transformer-based language models such as BERT (Devlin et al., 2019) and its successors RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019). For example, Wiedemann et al. (2020), the top performing team for the offensive language detection task at OffensEval 2020, use tweets from SOLID to fine-tune the masked language modeling objective of BERT-like models before training them on the labeled OLID data for the task of offensive language

detection. Other systems (e.g. Arslan (2020) for Turkish) rely on more tailored sets of features such as existing lexicons of offensive words.

In spite of the growing amount of work on this topic, little attention has been devoted to more sophisticated uses of lexical features and on the role of the time dimension in offensive language phenomena. Languages are subject to constant change and their lexicons are no exception. Over time, words can acquire new meanings, or change or lose existing ones (Koptjevskaja-Tamm, 2002). These changes in the semantic profile of a word can take various shapes. For example, they can widen or narrow its semantic scope or change its polarity. An example is *sick*, which gained a positive connotation of 'excellent, impressive; risky' (OED Online) in slang contexts in the early 1980s. Lexical semantic change is a highly complex phenomenon and its study helps us better understand the relation between language and social, cultural and historical factors, and how this relation changes over time. This has important consequences for all computational systems that rely on word lists as input, including those used in offensive language detection systems, as such lists tend to be static and therefore do not account for the changes that words are subject to.

This paper focuses on the task of offensive language detection and follows the framework set up in the Offenseval 2019 and 2020 competitions, particularly subtask A. We build on SINAI (Plaza-del Arco et al., 2019), the only lexicon-based system submitted to Offenseval 2019 for which we could access the code. Our system relies on a set of refined lexical features which cover the surface-level spelling of offensive content as well as their semantic change over a short time period. Our system is aware of the change in meaning of the word *Karen*, for example, which, according to Dictionary.com, in 2020 acquired an offensive meaning.[1] Our work shows that accounting for language change and more sophisticated lexical features can help research in offensive language detection. At the same time, we show that detecting semantic changes that occurred in a very short time interval (one year in our case) presents challenges because this phenomenon affects a small number of words. Focusing on offensive language (and therefore on words whose semantics changed towards or away from an offensive meaning) presents additional challenges, as this phenomenon affects an even smaller number of words. Our results are promising, especially considering that they are obtained by drawing on lexical features from datasets covering only two consecutive years, the only years for which the Offenseval training and test sets are available. During this period, only a small number of words changed their meaning or polarity. Therefore, we expect our method to have a bigger impact when tested on a larger time span. We stress that one important methodological strength of our method is that it does not need an up-to-date training set. Because it uses an older annotated dataset as its training set, it enables significant savings in the human effort and computational resources needed to create high-quality labelled data, while still being able to handle the ever-evolving lexical semantics of offensive language.

In addition to presenting a manually curated list of words that acquired or lost an offensive meaning between 2019 and 2020, we perform an extensive error analysis to explore the categories of texts that are misclassified by our system. We find that further improvements will likely come from better contextual representations: these will help prevent cases in which models detect offense in any text that contains a word that *might* be offensive in certain contexts. We suggest expanding the current set of offensive words will help to correctly label cases in which rare but offensive words are used.

2 Related work

2.1 Offensive language detection

There has been extensive work on offensive language detection, with a particular focus on platforms such as Twitter (Davidson et al., 2017; Lee et al., 2018; Founta et al., 2018); Reddit (Mittos et al., 2020; Hada et al., 2021; Ribeiro and Silva, 2019) and YouTube (Ottoni et al., 2018). Researchers have experimented with a range of classification models that strive to identify offensive language automatically. Early work relied on established machine learning techniques such as logistic regression (Waseem and Hovy, 2016) and SVMs (Karan and Šnajder, 2018). Researchers have also developed deep learning models to detect abusive language, e.g., using Convolutional Neural networks (Gambäck and Sikdar, 2017; Ribeiro

[1] "Karen is a pejorative slang term for an obnoxious, angry, entitled, and often racist middle-aged white woman who uses her privilege to get her way or police other people's behaviors" https://www.dictionary.com/e/slang/karen/.

and Silva, 2019), Gated Recurrent Unit networks (Zhang et al., 2018) and Long short-term memory models (Badjatiya et al., 2017), as well as ensemble architectures of neural with non-neural models (Anand et al., 2022). Recent large pre-trained language models have led researchers to experiment with transfer learning approaches (El-Alami et al., 2022; Guest et al., 2021; Sohn and Lee, 2019; Polignano et al., 2019). For example, Mozafari et al. (2019) fine-tune a pre-trained BERT model for hate speech detection, gaining F1-score of 88% and 92% on two datasets (Waseem et al., 2017; Davidson et al., 2017). The efficacy of BERT-based techniques has been evidenced in various competitions (Zampieri et al., 2019b, 2020).

One issue of the aforementioned classification approaches is that they need large and up-to-date annotated datasets for model training. To address this issue in offensive language detection, Singh and Li (2021) propose a domain adaptation training for bidirectional transformers to enhance the detection performance on a target dataset by exploiting an external dataset. However, this approach has three main limitations: 1) target and auxiliary datasets might not share the same label space, resulting in ad-hoc data transformations; 2) an external large-scale dataset relevant to the target task is still needed; and 3) adding time-independent information does not remove the need for up-to-date annotated datasets. Indeed, language is subject to constant change: new words emerge all the time to refer to new concepts, for example, and existing words acquire new meanings (or lose their existing ones), a phenomenon that affects mainly open-class items (nouns, verbs, adjectives and adverbs), including offensive terms. By introducing a semantic change module, our work leverages time-dependent lexical features for offensive language detection which, in turn, could lighten the burden on having large-scale and up-to-date data.

2.2 Lexical semantic change detection

Over the past 15 years, the area of lexical semantic change detection has attracted a growing level of attention (Tahmasebi et al., 2018; Kutuzov et al., 2022). This task aims at identifying which words changed their meaning in a given time period. Researchers have proposed a range of methods to address this, from graph-based models (Mitra et al., 2015; Tahmasebi and Risse, 2017) to topic models (Cook et al., 2014; Lau et al., 2014; Frermann and Lapata, 2016), but the most successful methods involve type or token word embeddings (Kim et al., 2014; Basile and McGillivray, 2018; Kulkarni et al., 2015; Hamilton et al., 2016; Dubossarsky et al., 2017; Tahmasebi, 2018; Rudolph and Blei, 2018; Jatowt et al., 2018; Tang, 2018).

The most common approach to lexical semantic change detection consists in building type or token embedding representations of the semantics of words from an input diachronic corpus, which is split into subcorpora covering different time intervals. If type embeddings are used, these need to be aligned over the temporal sub-corpora, usually via orthogonal Procrustes (Hamilton et al., 2016), vector initialisation (Kim et al., 2014) or temporal referencing (Dubossarsky et al., 2019). Finally, significant shifts which can be interpreted as indications of semantic change are detected by measuring the change between the representations of the same word over time. This is typically done via distance metrics based on cosine or local neighbours.

In 2020 the first standard evaluation framework and dataset for this task were created for the SemEval 2020 shared task on Unsupervised lexical semantic change detection (Schlechtweg et al., 2020). The best-performing systems in this task use type embedding models, although the quality of the results differs depending on the language. Averaging over all four languages, the best result had an accuracy of 0.687 for sub-task 1 and a Spearman correlation coefficient of 0.527 for sub-task 2.

3 Approach

3.1 Overview

We experimented with enriching a lexicon-based offensive language detection system (OLD) with time-sensitive lexical features derived from lexical semantic change detection (LSCD) in a new system which we called LSCD+OLD. The idea behind this is to rely on "dated" manually annotated data to train a classifier that can label new instances of text for its offensiveness. Imagine that we have data annotated in 2019 (e.g. the OLID dataset) and we are interested in detecting offensive language in 2020. We expect that most of the words have not changed their meaning between 2019 and 2020, but some have, and a portion of those have acquired (or lost) an offensive meaning. These new meanings will not be recorded in the 2019 data and therefore our classifier is likely to miss instances of offensive texts if they contain one or

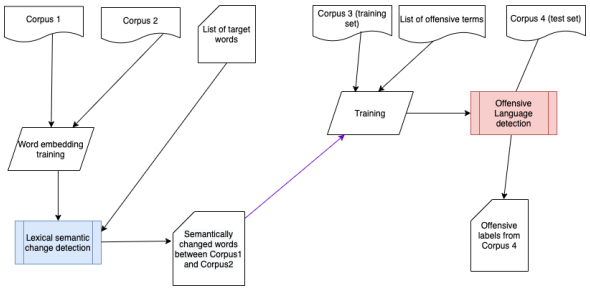

Figure 1: Diagram of our system, with the Semantic Change Detection (left) and the Offensive Language Detection (right) modules.

more of these words. We propose to overcome this by incorporating semantic change knowledge into the system. This means that we would not need to engage in the expensive process of producing new manually annotated data from 2020. Figure 1 shows the architecture of our system and the next sections describe its modules. The two modules are joined by a series of features that use the output of the LSCD module as input to the OLD classifier. Our code is available at https://github.com/alan-turing-institute/offenseval-semantic-change.

3.2 Lexical semantic change detection module

The *LSCD module* (left hand side of Figure 1) takes as input two corpora, representing the first time period t_1 (typically the time period when the manually annotated dataset was produced, 2019 in our case) and the second time period t_2 (the time after the manually annotated dataset was produced, 2020 in our case), respectively. Word embeddings are trained on the two corpora. For every target word in the intersection between the two vocabularies, the LSCD module outputs a semantic change score, representing the degree by which the word has changed between t_1 and t_2.

We chose UWB (Pražák et al., 2020) for the implementation of the LSCD module. UWB ranked first in sub-task 1 of SemEval 2020's shared task on unsupervised lexical semantic change detection, with an absolute accuracy of 0.687, which was the best result on average over all four languages (English, German, Latin, and Swedish). UWB involves training word embeddings for each of the two time-separated corpora, setting up two semantic vector spaces. Canonical Correlation Analysis, using the implementation from Brychcin et al. (2019) and a modification of the Orthogonal Transformation from VecMap (Artetxe et al., 2018) are then used to compute a linear transformation between the earlier and later spaces. Finally, the cosine distance between the transformed vector for the target word from the earlier corpus and the vector for the target word in the later corpus is measured and presented as the semantic change score. UWB's system consists of the following adjustable hyperparameters, with which we experimented with: (i) *Embedding dimensions*: the dimensions of the continuous vector space onto which the learned word representations are translated; (ii) *Window size*: the number of adjacent words used to determine the context of each word; (iii) *Iterations*: the number of times parameters are updated; (iv) *Minimum frequency count*: the minimum frequency below which uncommon words are set to unknown; and (v) *Maximum links*: the maximum number of links, i.e. size of vocabulary.

3.3 Offensive language detection module

The lexicon-based *OLD module* takes as input a training set and a list of offensive terms, which are used to train a classifier that can label a new set of texts as offensive or not. A description of the datasets is given in Section 4. The offensive language detection component of the proposed system is based on SINAI, developed by Plaza-del Arco et al. (2019) for SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). SINAI uses OLID data for training and testing and was the only lexicon-based system for which we could find the underlying code and were able to reproduce the Offenseval 2019 results. SINAI was chosen because the description of the other lexicon-based systems available in the corresponding system description papers for the Offenseval shared task were not sufficiently detailed to ensure that our implementation would have led to the same results as the original systems.

The system preprocesses the OLID data to remove mentions and URLs, and tokenize the tweets. It then trains a Support Vector Machine (SVM) classifier on *statistical features* (specifically TF-IDF scores) and the following two *lexical features*. (1) *Sentiment*: vaderSentiment[2] is used to obtain a vector with four scores: negative, positive, neutral, and compound polarity; and (2) *Offensive word list*: the proportion of tokens in the Offensive/Profane Word List[3] out of all tokens in each

[2] https://pypi.org/project/vaderSentiment/

[3] https://www.cs.cmu.edu/~biglou/

Table 1: Summary of OLID and SOLID datasets

Dataset	Tweets in training set	Tweets in test set
OLID	13,240	860
SOLID	6,209,964	3,887

tweet. In addition to SINAI's original features, we introduce three additional lexical features, a time-independent one and two time-dependent ones. The effect of introducing the time-dependent ones is that they consider the semantic change that affected words between 2019 and 2020 and therefore allow the classifier to "update" the 2019 training set.

Character length: For each tweet, the average number of characters of its tokens if they are contained in a offensive word list; this is based on the fact that many highly offensive words in English are short (typically four characters) (Bergen, 2016).

Polarity change: for every token found in both the 2019 and the 2020 corpora, we calculate the proportion of negative-sentiment tweets the token occurred in out of all tweets it occurred in. We calculate the difference between these two proportions and divide it by the 2019 proportion, to obtain the token's rate of change in proportion of negative tweets over time. For every tweet, we take the maximum polarity change value of all its tokens.

Sentiment and semantic change scores: for each token we multiply its semantic change score by its polarity change score as defined in the Polarity change features, and take the maximum value. This way, we aim to capture those words which underwent usage change and polarity change combined, with the idea to approximate the detection of those words that not only changed semantically, but whose semantics changed in an offensive direction.

4 Data

We rely on the OLID training dataset, which includes 13,240 tweets from late 2018 and 2019 annotated according to a three-layer hierarchical annotation scheme. The first layer identifies a tweet as containing offensive language (OFF) or not (NOT). The second layer categorizes the offensive language in tweets as a targeted insult (TIN) or an untargeted insult (UNT). The third layer categorises the targets of insults as an individual (IND), a group (GRP), or other (OTH) (Zampieri et al., 2019a). The OLID test set includes tweets categorized according to the sub-tasks, along with their gold labels. The offensive language detection system of our model uses the OLID training set for sub-task A.

SOLID contains tweet IDs for over 9,000,000 tweets from early 2020, also annotated according to the three-level hierarchy of OLID. We extract the content of over 6,000,000 tweets using the Twitter API by matching the SOLID tweet IDs. Contrary to the OLID dataset, SOLID does not contain gold standard labels for any of the sub-tasks. Instead, SOLID uses a democratic co-training method to provide the average confidence (AVG_CONF) and standard deviation from the AVG_CONF (CONF_STD) values of a particular tweet belonging to the positive class of that sub-task. For sub-task A, a tweet belongs to the positive class if it is labelled as offensive, or OFF (Rosenthal et al., 2020). We utilise a similar method as that used by Plaza del Arco et al. (2020) to generate the tweet labels using the AVG_CONF and CONF_STD values. We take 0.5 to be the threshold value for a tweet to be labelled offensive. If, for a given tweet, the AVG_CONF + CONF_STD value is still below the threshold value of 0.5, we label the tweet as NOT. If the AVG_CONF - CONF_STD gives a value more than the threshold, we label the tweet as OFF. Any tweets whose AVG_CONF + CONF_STD values were greater than 0.5, or AVG_CONF - CONF_STD values were less than 0.5 are discarded, as this indicates the OFF/NOT classification is not strongly established, and varies based on the standard deviation.

Table 1 gives a summary of the OLID and SOLID datasets. Following Plaza-del Arco et al. (2019), we preprocess the OLID and SOLID datasets by tokenizing the tweets using NLTK, lower-casing all tokens and removing URLs and Twitter user mentions.

4.1 Twitter corpora

In order to collect Twitter data from several years in the past, we download samples collected by the Archive Team.[4] These samples are taken from the Twitter 1% streaming API from 2012 until the time of the present study. We use only a small portion of this dataset from each year in order to keep the training time of the semantic change model manageable. We select a sample from the same time of each year (beginning of March). We obtain an average of 114,995 tweets for each year. More

resources/

[4] https://archive.org/details/twitterstream

Table 2: Summary of Twitter data sample to be used for LSCD module.

Year	Tweets	Tokens
2019	226,275	2,624,412
2012-2019	919,965	8,391,550
2020	364,708	4,205,419

statistics about the dataset can be found in Table 2.

We remove URLs, Twitter handles, and punctuation marks, and apply lower-casing. We correct cases in which the same character is repeated in a string (e.g. *faaast* vs. *fast*). We also lemmatise the text and exclude strings with fewer than three characters, with the exception of a fixed list of function words like pronouns and prepositions. Finally, we replace emoji with corresponding text using the emoji[5] Python package and tokenize using the Twitter tokenizer in NLTK (Bird et al., 2009). This last step was taken to simplify the data processing. However, we recognise that replacing emoji with their names will not capture the semantic change of emoji themselves, which we have investigated in one of our previous studies (Robertson et al., 2021). In future work we could look into incorporating these changes into our system.

4.2 Ground truth for semantic change

We compile a list of words for the evaluation of the LSCD module. These words not only changed their semantics between 2019 and 2020 but also did so by acquiring a new offensive meaning. We analyse a mix of online sources in order to identify offensive words whose definitions shifted between 2011 and 2019: Hatebase, an online repository of words associated with hate speech, and earlier academic offensive language lists, namely those by Luis von Ahn (Horta Ribeiro et al., 2018),[6] and by ElSherief et al. (2018b,a).[7]

We search Urban Dictionary[8] and `Dictionary.com` to confirm definitions and dates of meaning change. The criterion for selecting words from previously compiled lists was that they had to display at least one non-offensive and one offensive definition. We rely on Urban Dictionary, a crowd-sourced slang language dictionary, to verify definitions and to approximate the date at which a change occurred. We then search Dictionary.com's slang definition list of almost 1,000 words and phrases to find new negative connotations to existing words.

Lexical semantic change over a short time period such as the one considered here is a low-frequency phenomenon. Moreover, lexical semantic change involving a new offensive sense, which emerges alongside the established non-offensive senses, is an even rarer phenomenon. For this reason, the list had to be further refined to make sure that the corpora at our disposal displayed evidence of the words having undergone this phenomenon.

For example, *beta* occurs 49 times in the 2019 Twitter corpus and 67 times in the 2020 Twitter corpus. In 2019 the majority of its usages refer to the neutral software-related meaning reported by the Oxford English Dictionary as "a test of machinery, software, etc. in course of final development, carried out by a party or parties unconnected with the developer" (bet, 2021) as in (1) and none of the 2019 usages are offensive. On the other hand, the 2020 data show eight offensive usages out of 56 of "a slang insult for or describing a man who is seen as passive, subservient, weak, and effeminate",[9] as in (2):

(1) RTL Release 0.2.16-beta New Feature: Routing Peers - Routing history analysis by Peers (requested by @USER)

(2) Jelly viagra these man r so beta

For each of the selected words, we conduct a diachronic corpus analysis to check that the word was used in an offensive sense more often in the 2020 corpus than in the 2019 corpus. Through this, we obtain a subset of 21 lemmas. For each of these 21 lemmas we search for a corresponding stable lemma which did not acquire a new sense in 2020 (as checked against the Oxford English Dictionary) and which had similar frequency counts in the 2019 and 2020 corpus and same part of speech. In Appendix A, Table 7 shows the final list of 42 lemmas and Table 8 shows the list of positive gold standard words, i.e. the words that acquired an offensive meaning. The gold standard words are: *beta, canceled, cap, cringe, fag, globalist, karen, monkey, mug, ratchet, salty, simp, skip, snowflake, sus, thirsty, illegal, chad, gammon, Brexiter, triggered*. Appendix A contains additional information about their semantics.

[5] https://pypi.org/project/emoji/

[6] https://github.com/manoelhortaribeiro/HatefulUsersTwitter/blob/master/data/extra/bad_words.txt

[7] https://github.com/mayelsherif/hate_speech_icwsm18/blob/master/hate_keywords.txt

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

[9] https://www.dictionary.com/e/slang/beta/

5 Experiments

In this section, we present the experiments performed to find the best configuration of parameters for our system.

5.1 Experimental setup

Our aim is to assess whether it is possible to train an OLD system on older data and achieve comparable performance when using this system to classify newer data. Therefore, we use the OLID training dataset as our training set, and the SOLID test set as our test set. During the development phase we could not use a portion of the OLID training set because its content is not from the same time period covered by SOLID. Therefore, we use a portion of SOLID as development — 0.2% of its data, corresponding to 9,915 tweets. We train the linear SVM classification algorithm (SVC) with C parameter 0.5 on SINAI's original features and also experiment with the additional lexical features we introduced in section 3.3.

Semantic change detection As part of the LSCD module of our system, we run the UWB code on two sets of corpora: Twitter 2019 vs. Twitter 2020; and Twitter 2012-2019 vs. Twitter 2020. Even though the time periods covered by OLID and SOLID are 2019 and 2020, respectively, we also want to see whether expanding the time period further back helps the performance.

Word embedding training As an input, we train word type embeddings with the following parameters: embedding dimensions: 100, 300, 1000; window size: 2, 5, 10; iterations: 5; min freq count: 1, 5, 10, 50, 100; embedding type: fasttext and word2vec; and embedding algorithm: skipgram and continuous-bags-of-words

Change detection We run the UWB code for the LSCD module of our system. We then train the SVC classifier on the features listed in section 3.3, setting three values for the threshold on the semantic change score: 0.5, 0.7, and 0.9.

5.2 Results

Our best model uses von Ahn's list of offensive words, the features described in Section 3.3 plus the TF-IDF features and SINAI's lexical featurxes, but not SINAI's sentiment features. The best system is based on the following parameters for the LSCD module: t_1 = 2019 and t_2 = 2020, word2vec embeddings with the continuous bag of words algorithm, 1000 dimensions, 5 iterations, a context window of 10, a minimum frequency count of 10 and 100000 maximum number of links used by the UWB code.

Table 3: Evaluation results of the lexical semantic change module against our gold standard by different threshold values applied to the semantic change scores. The fifth row shows the number of words in the positive gold standard set that were also found as positive candidates for semantic change.

Metric	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
F1	0.93	0.93	0.57	0.42	0.24	0.24
Acc	0.89	0.92	0.64	0.58	0.50	0.50
Prec	1.00	1.00	1.00	1.00	1.00	1.00
Rec	0.87	0.87	0.40	0.27	0.13	0.13
#GS words	12	12	6	4	2	0

Table 4: Results of the quantitative comparison between the lexical semantic change output and the positive gold standard words.

Metric	**Value**
average score (positive gold standard)	0.60
median score (positive gold standard)	0.59
average score (other words)	0.49
median score (other words)	0.52
Mann-Whitney statistic	87631
Mann-Whitney p-value	0.02

The output of the LSCD code is the list of the vocabulary words, paired with a lexical semantic change score. The higher the score the higher the likelihood that the word underwent lexical semantic change. In order to obtain a list of candidates for semantic change, a threshold must be set for the score: all words with a score above the threshold are then considered as candidates. We evaluate the LSCD module against the gold standard described in Section 4.2. True positives (TP) are the words that are identified as having undergone semantic change and that also appear in the gold standard set of changed words. True negatives (TN) are the words that are identified as not having changed and also appear in the gold standard set of unchanged words. False positives (FP) are the words that are identified as having changed but are in the gold standard list of unchanged words. False negatives (FN) are the words that are identified as not having changed but are in the gold standard list of changed words. Accuracy is calculated as $(TP + TN)/(TP + FP + TN + FN)$. Precision is calculated as $TP/(TP + FP)$ and recall as $TP/(TP + FN)$.

Table 3 shows the results. Table 4 shows an

Table 5: Results of experiments on the OffensEval 2020 test set; all systems were trained on the OLID training set (2019), apart from the last one, which was trained on SOLID (2020).

	SINAI	**LSCD+OLD**	**SINAI**
Training Data	*OLID*	*OLID*	SOLID
Test Data	SOLID	SOLID	SOLID
Precision (NOT)	0.97	0.97	**0.99**
Recall (NOT)	0.91	**0.92**	0.90
Prec. (OFF)	0.79	**0.82**	0.79
Recall (OFF)	0.93	0.93	**0.98**
Prec. (Macro)	0.88	**0.90**	0.89
Prec. (Weighted)	0.92	0.93	**0.94**
Recall (Macro)	0.92	0.93	**0.94**
Recall (Weighted)	0.91	**0.92**	**0.92**
Accuracy	0.91	**0.92**	**0.92**
F1 (Macro)	0.90	**0.91**	**0.91**

additional analysis aimed at measuring the output against the gold standard by comparing the average and median lexical semantic change score of the positive gold standard words and of the other words. Tables 3 and 4 show that the algorithm's performance with a threshold of 0.5 (the threshold chosen for the final model) is very good, even better than the current state-of-the-art from the SemEval 2020 task 1 results, where UWB achieved an average accuracy of 0.687 on the four languages and 0.622 on English. The setup of that shared task was very different to this study, as the English dataset covered a much longer time span (t_1 =1810–1860 and t_2 =1960–2010). Table 4 shows that the set of positive gold standard words have a significantly higher semantic change score compared with the other words, with an average of 0.60 vs. 0.49 and a median of 0.59 vs. 0.52, respectively.

Table 5 shows how our system compares to the original SINAI system trained on OLID and on its version trained on SOLID. The three system's performances are generally quite close to each other, with small differences. Our system combining extra general and time-dependent lexical features into SINAI performs better than the baseline in all metrics apart from the precision on the NOT class where it achieves the same results as the baseline.

It is interesting to note that our system achieved a macro-averaged F1 score of 0.94 on the development set drawn from 0.02% of the SOLID training set. This result may be explained by the fact that a larger set is more likely to capture a higher number of words that acquired an offensive meaning between 2019 and 2020, since this is a low-frequency phenomenon as we have seen. This suggests that our system may achieve even better performance when tested against a larger time span than the one-year period studied here.

6 Error analysis

In order to gain a better understanding of the 297 errors made by our proposed system, we qualitatively inspected the misclassified examples and sorted them into seven major categories (summarized in Table 6). Each misclassified instance was categorized by two of the authors. We calculated the Inter-Annotator Agreement (IAA) as Cohen's $\kappa = \frac{\sum a - \sum ef}{N - \sum ef}$, where $\sum a$ is the number of agreements, $\sum ef$ is the sum of the expected frequencies of agreement by chance, and N is the the number of misclassified instances (Cohen, 1960). We interpreted the IAA scores according to the following criteria: 0.01-0.20 points to no agreement/ slight agreement, 0.21-0.40 to fair agreement, 0.41-0.60 to moderate agreement, 0.61-0.80 to substantial agreement, and 0.81-1.00 to strong agreement. The average of the pairwise agreement is moderate (0.46), see Table 9 in Appendix A. This shows that the task is quite difficult, even for human annotators.

To present the analysis of the distribution of the seven categories we identified in the list of errors, we focus on 178 errors whose classification the two annotators agreed on. For examples that were misclassifed as offensive, the most common feature was the presence of offense-related words that were not being used in an offensive way. Models based solely on lexical features that do not account for the contextual meaning of words will naturally struggle with these cases, and the high number of these errors suggests that improving the semantic change detection may not have as large of an impact compared to including better contextual representations of potentially offensive words. The next most common category of misclassified offensive was self-deprecating statements: those that used statements that *would* be considered offensive had they been targeted at someone else, but they were instead directed at the author of the text (e.g., "I am ugly"). A few statements misclassified as offensive actually employed irony, in which the surface meaning of the text appears offensive, but the intended meaning of the text was not offensive.

For texts misclassified as not offensive, the most commonly noticed feature was that an offensive word was present, but it was not included in the offensive word list that was used by the best performing model. Some of these words acquired

Table 6: Analysis of error categories. The columns represent the category of error agreed upon by two annotators; the second and third column contain the incorrect prediction by our model and the last column contains the total counts.

Error category	NOT	OFF	Total
offensive-related words not used in an offensive way	0	96	96
offensive word not in list	20	0	20
self-deprecating	1	18	19
indirect offense	10	6	16
incorrect groundtruth	9	3	12
other/unexplained	5	7	12
irony	0	2	2

an offensive meaning over time. An example is "@USER I thought you magas refused to use Nike because they don't hate black people". In this tweet, the word *magas* is a case of a semantically-changed word which here is employed in an offensive way. In Urban Dictionary, this word is defined in November 2018 as "a word used in the campaign of trump, signals neo nazis and white supremacists". For these examples, we hypothesize that better expansion of the offensive word list may help with being able to correctly categorize these examples. For both types of misclassification, a handful of the instances contained more indirect examples of offense, which has been highlighted as an important category to focus on within the offense detection domain that may require multi-hop reasoning (Zhang et al., 2022). A small number of other examples appear to have been incorrectly labeled in the original dataset, and a handful were difficult to categorize (other) or understand (unexplained).

7 Conclusion

We have presented a study on combining lexical semantic change information into a new system that performs offensive language detection based on lexical features, and a curated gold standard list of English words that acquired or lost an offensive meaning between 2019 and 2020. From the point of view of the performance, our system trained on the much smaller and older OLID data performs better than SINAI trained on the same data. Further, by including our time-dependent lexical features, our system, trained only on the older OLID data, has performance on the newer SOLID test set that is comparable to a SINAI model that was trained directly on the much larger and newer SOLID training set. This shows that it indeed language change affects offensive language and it is possible to perform offensive language detection by taking into account such change and without relying on large labelled datasets that have been produced around the same time as the texts on which it is applied. Additionally, we discuss the challenges of performing short-term semantic change detection, especially for the rare words that acquired or lost an offensive meaning over a period of two years. Future work will involve expanding our evaluation across other time periods and corpora.

8 Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This research was supported in part through computational resources provided by The Alan Turing Institute and with the help of a generous gift from Microsoft Corporation. The work of Chiara Di Bonaventura was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). We would like to thank Dr Gard Jenset for his advice on the definition of the features and the study design and Madeline Tondi for curating the gold standard word list.

9 Author contributions

BMcG designed the study, managed and supervised the project. She also carried out the evaluation of the lexical semantic change module, the experiments with the offensive language detection module, refined the gold standard list and wrote Sections 1, 2.2, 3-3.3, 4.1, 5-5.2, and 6. JC reproduced the code for semantic change detection, contributed to the study design and wrote section 3.3. MA reproduced the code for offensive language detection and wrote Section 4. GT contributed to the design of the study and the supervision of the project; he wrote Section 2.1 and contributed to writing across the remaining sections. SW contributed to the design of the study and the supervision of the project, wrote part of Section 4.1, and helped edit the other sections. CDB contributed to Sections 2.1 and 6. AMP contributed to Sections 1 and 6. BMcG, MA, CDB, AMP, and SW performed the annotations used in section 6.

References

2021. "beta, n.". OED Online, Oxford University Press.

M Anand, Kishan Bhushan Sahay, Mohammed Altaf Ahmed, Daniyar Sultan, Radha Raman Chandan, and Bharat Singh. 2022. Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. *Theoretical Computer Science*.

Pinar Arslan. 2020. Pin_cod_ at SemEval-2020 task 12: Injecting lexicons into bidirectional long short-term memory networks to detect Turkish offensive tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2117–2122, Barcelona (online). International Committee for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Pierpaolo Basile and Barbara McGillivray. 2018. *Discovery Science*, volume 11198 of *Lecture Notes in Computer Science*, chapter Exploiting the Web for Semantic Change Detection. Springer-Verlag.

Benjamin K. Bergen. 2016. The Science of Swear Words (Warning: NSFW AF). https://www.wired.com/2016/09/science-swear-words-warning-nsfw-af/. Accessed: 2021-08-26.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tomas Brychcin, Stephen Eugene Taylor, and Lukas Svoboda. 2019. Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Syst. Appl.*, 135:287–295.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (1):37–46.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

Fatima-zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6048–6056.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Y. Wang, and Elizabeth Belding. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, ICWSM '18.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018b. Peer to peer hate: Hate instigators and their targets. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, ICWSM '18.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for english reddit comments. *arXiv preprint arXiv:2106.05664*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1489–1501.

Manoel Horta Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Adam Jatowt, Ricardo Campos, Sourav S Bhowmick, Nina Tahmasebi, and Antoine Doucet. 2018. Every word has its history: Interactive exploration and visualization of word sense evolution. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1899–1902. ACM.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd workshop on Abusive Language Online (ALW2)*, pages 132–137.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *LTCSS at ACL*, pages 61–65.

M. Koptjevskaja-Tamm. 2002. The lexical typology of semantic shifts.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized embeddings for semantic change detection: Lessons learned. *Northern European Journal of Language Technology*, 8(1).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 259–270.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. *arXiv preprint arXiv:1808.10245*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. 2020. "and we will fight for our race!" a measurement study of genetic testing conversations on reddit and 4chan. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 452–463.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940.

Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernardina, Wagner Meira Jr, and Virgílio Almeida. 2018. Analyzing right-wing youtube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM conference on web science*, pages 323–332.

Flor Miriam Plaza-del Arco, M. Dolores Molina-González, Maite Martin, and L. Alfonso Ureña-López. 2019. SINAI at SemEval-2019 task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 735–738, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Flor Miriam Plaza del Arco, M. Dolores Molina González, Alfonso Ureña-López, and Maite Martin. 2020. SINAI at SemEval-2020 task

12: Offensive language identification exploring transfer learning models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1622–1627, Barcelona (online). International Committee for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, and Giovanni Semeraro. 2019. Hate speech detection through alberto italian language understanding model. In *NL4AI@ AI* IA*.

Ondřej Pražák, Pavel Přibáň, Stephen Taylor, and Jakub Sido. 2020. UWB at SemEval-2020 task 1: Lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online). International Committee for Computational Linguistics.

Alison Ribeiro and Nádia Silva. 2019. Inf-hateval at semeval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425.

Alexander Robertson, Farhana Ferdousi Liza, Dong Nguyen, Barbara McGillivray, and Scott A. Hale. 2021. Semantic journeys: Quantifying change in emoji meaning from 2012–2018. In *Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media*, online.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. page 14.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Sumer Singh and Sheng Li. 2021. Exploiting auxiliary data for offensive language detection with bidirectional transformers. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 1–5, Online. Association for Computational Linguistics.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.

Nina Tahmasebi. 2018. A study on word2vec on a historical Swedish newspaper corpus. In *CEUR Workshop Proceedings. Vol. 2084. Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, Helsinki Finland, March 7-9, 2018.*, Helsinki. University of Helsinki, Faculty of Arts.

Nina Tahmasebi, L. Borin, and A. Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv: Computation and Language*.

Nina Tahmasebi and Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *arXiv preprint arXiv:1705.09899*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644, Barcelona (online). International Committee for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. pages 1415–1420.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2022. Rethinking offensive text detection as a multi-hop reasoning problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3888–3905, Dublin, Ireland. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

A Appendix

Table 7: Gold standard list of words that acquired an offensive sense for which there is evidence in our 2020 corpus (left) and stable words (right). For each group, the first column provides the part of speech, the second the lemma, the third the frequency in the 2019 Twitter corpus and the fourth the frequency in the 2020 Twitter corpus.

pos	**Lemma**	**Freq 2019**	**Freq 2020**	**Stable word**	**Freq 2019**	**Freq 2020**
N	beta	49	67	academy	50	65
ADJ	cancel	25	64	acceptable	23	63
N	cap	153	221	fish	151	228
ADJ	cringe	36	65	additional	33	72
N	fag	15	20	accuracy	16	17
N	globalist	17	21	absurd	23	23
N	karen	39	106	behaviour	48	99
N	monkey	60	97	corporation	59	99
N	mug	68	93	cage	62	85
N	ratchet	2	12	moonstone	2	3
ADJ	salty	17	31	alcoholic	19	22
N	simp	5	67	whorehouse	5	63
N	skip	155	148	abandonment	5	10
N	snowflake	11	27	calamity	2	19
ADJ	sus	11	24	beneficial	8	20
ADJ	thirsty	33	35	dreamy	26	29
N	illegal	170	217	direction	163	223
N	chad	8	23	contestant	9	213
N	gammon	4	6	gravel	3	8
N	Brexiter	1	9	grenade	3	12
ADJ	triggered	31	31	analytic	27	39

Table 8: Gold standard list of words that acquired an offensive meaning, the date of its first recorded usage and the source dictionary.

pos	Lemma	Offensive meaning	Source	Date
N	beta	Insult describing a man who is seen as passive, subservient, weak, and effeminate.	https://www.dictionary.com/e/slang/beta/	1990
ADJ	canceled	When a person is canceled, they are no longer supported publicly. Sometimes used as a threat, "to cancel."	https://www.dictionary.com/e/pop-culture/cancel-culture/	2015
N	cap	A lie.	https://www.urbandictionary.com/define.php?term=cap	2020
ADJ	cringe	Someone or something extremely embarrassing or awkward.	https://www.urbandictionary.com/define.php?term=Cringe	2013
N	fag	A derogatory term for homosexual.	https://www.urbandictionary.com/define.php?term=fag	2010
N	globalist	Coded language often used as a negative euphemism for Jew.	https://www.urbandictionary.com/define.php?term=Globalist	2018
N	karen	Karen is a pejorative slang term for an obnoxious, angry, entitled, and often racist middle-aged white woman who uses her privilege to get her way or police other people's behaviors.	https://www.dictionary.com/e/slang/karen/	2020
N	monkey	A derogatory term for a black person.	https://www.urbandictionary.com/define.php?term=Monkey	2011
N	mug	Unattractive, unappealing, or unpleasant.	https://oed.com/view/Entry/89666161?rskey=Vq0ZKB&result=0&isAdvanced=true#firstMatch	2009
N	ratchet	Someone whose actions could be considered as severely undistinguishable; possesing little or no class.	https://oed.com/view/Entry/89666161?rskey=Vq0ZKB&result=0&isAdvanced=true#firstMatch	2009
ADJ	salty	Angry, upset, or hostile, especially due to embarrassment or failure.	https://www.dictionary.com/browse/salty	2011
N	simp	Simp is a slang insult for men who are seen as too attentive and submissive to women, especially out of a failed hope of winning some entitled sexual attention or activity from them. Can also refer to an avid fan of a celebrity.	https://www.dictionary.com/e/slang/simp/	2011
N	skip	A white Australian, alluding to Skippy the Bush Kangaroo, a once-popular Australian television show for children.	https://www.dictionary.com/browse/salty	2011
N	snowflake	A political insult for someone who is perceived as too sensitive, often used against young people and those with progressive political viewpoints.	https://www.dictionary.com/browse/snowflake	2015
ADJ	sus	Giving the impression that something is questionable or dishonest; suspicious.	https://www.dictionary.com/browse/snowflake	2015
ADJ	thirsty	Describes a graceless need for approval, affection or attention, to the point of another becoming uncomfortable.	https://www.dictionary.com/browse/snowflake	2015
N	illegal	Derogatory term for a Hispanic or Latino person in the United States.	https://www.dictionary.com/browse/snowflake	2015
N	chad	A rude, and often sexually promiscuous, man.	https://www.urbandictionary.com/define.php?term=Chad&page=4	2017
N	gammon	A term used against anyone who was white and voted for Brexit.	https://www.urbandictionary.com/define.php?term=Gammon	2018
N	Brexiter	An derogatory term to refer to someone who voted for Brexit.	https://www.urbandictionary.com/define.php?term=Brexiter	2016
ADJ	triggered	An emotional/psychological reaction caused by something that somehow relates to an upsetting time or happening in someone's life.	https://www.urbandictionary.com/define.php?term=Triggered	2016

Table 9: Pairwise Inter-Annotator Agreement (IAA) scores for the error analysis.

Annotators	**IAA**
A_1 & A_2	0.53 (moderate)
A_2 & A_3	0.88 (strong)
A_1 & A_3	0.09 (disagreement)
A_2 & A_4	0.39 (fair)
A_4 & A_5	0.40 (fair)
A_3 & A_5	0.37 (fair)
A_3 & A_4	0.56 (moderate)
A_1 & A_4	0.46 (moderate)