Argot as a Trust Signal: Slang, Jargon & Reputation on a Large Cybercrime Forum

Jack Hughes, Andrew Caines, and Alice Hutchings
Department of Computer Science and Technology
University of Cambridge
firstname.lastname@cl.cam.ac.uk

Abstract

We apply signalling theory to a cybercrime forum to explore how argot (slang and jargon) is used to signal trust in untrustworthy environments. We develop an argot detection tool, using word embeddings from forum and non-forum datasets, which are aligned using training annotations. Compared with prior work, our approach improves performance, with an increase in the F1 and accuracy scores. Using the detected argot to create per-user variables, we find a negative correlation between the use of argot and reputation votes. We explore the trajectories of groups of forum members to observe how the use of argot and user reputation in the forum varies over time. Our findings indicate forum users are using argot to overcome the cold start problem, a conundrum faced by new users to social networks with ranking systems and marketplaces with feedback systems. A significant group of long-standing users is characterised by high levels of argot in their early forum postings. This decreases once reputation metrics increase. This particular trajectory group are amongst the highest-rated long-term members.

1 Introduction

Signalling theory [1] has been used to explain how criminals in the real world use subtle clues to signal trustworthy aspects to others in an environment of low trust. One example of signalling in the underground is the use of tattoos. These can be used to demonstrate toughness and resilience to pain. Some tattoos signal the gang membership of the bearer and status within groups. Tattoos are difficult to remove, making them a permanent indicator of commitment [1]. Other groups use signals that are not so obvious, only caring to display group membership to others within that group, without bringing attention to themselves from outsiders. An example of a signal that relies on argot (slang and jargon) is Polari, a lexicon used by gay men during the time homosexuality was criminalised in the UK [2].

Signalling trustworthiness is particularly important for criminals, where they risk interacting with undercover police, and in environments where scruples are generally low. The idea of signalling theory is that signals of trustworthiness are cheap to emit–in this context, by authentic criminals (the way someone dresses, talks, or the tattoos they display on their bodies)–but expensive to mimic by those who are not genuine.

Signalling theory is particularly important when it comes to understanding how cybercrime forums operate [3, 4, 5, 6, 7, 8]. In these online spaces, you find people who need to interact with each other, to buy products or exchange services [8]. Anonymity often means they don't know the identity of their contact, and there is no threat of violence that might otherwise deter others from cheating. Nonetheless, the cost of interacting with the wrong person can be painful, including losing their money or their liberty. However, physical signals are almost completely absent. The main medium by which users interact with each other is by text. Therefore, we believe text-based signals are important for communicating trustworthiness.

To avoid becoming 'lemon markets' resigned to the control of rippers who price out genuine sellers [9], forums introduce reputation systems as a signal of trustworthiness. However, sybil attacks could be used to disrupt ('lemonise') the market, with false accounts promoting distrust within reputation systems [10]. Reputations systems provide a way for members to assign positive or negative votes to each other, and reputation metrics are displayed on members' profiles to inform other members. Reputation systems are not failsafe, in that they may be gamed to falsely gain a higher reputation score (or to attack a competitor). Another problem with reputation votes is how reputation is initially gained by those who are new and untrusted, when they require others to trust them to gain reputation votes. In economics, this conundrum is known as the *cold start problem*.

In this research into the cybercrime underground, we find the use of slang and jargon by members can signal a level of knowledge and trustworthiness to other members. We find that the use of argot is negatively correlated with reputation metrics. However, when we explore the trajectories of actors who have been active on the forum for a year or more, we stumble upon more nuanced results. We find a significant group of users whose use of argot drops, while their reputation increases. These users are amongst the most active and highly reputed. Initially we measure high levels of argot being used by this group. However, this rapidly decreases, just months into their forum activity, coinciding with a growth in reputation. We believe this

group of users is using argot to overcome the cold start problem. Once they become highly reputed, reputation metrics take over as a trust signal, and their use of argot diminishes, as using a specialised language is likely to take considerable cognitive effort to learn [2].

In this work, we use the word 'argot' to refer to the slang and jargon used by a particular group – in this case, cybercrime forum members. By 'slang' we mean colloquial language either of new words or current words used in a different sense, and by 'jargon' we mean words used by a specific group that are difficult for others to understand, such as technical terminology.

Let's consider the term 'rat'. The majority of people may immediately think of a rodent, typically bigger than a mouse, common in highly populated areas. However, within criminal communities, the term rat would be used in relation to someone untrustworthy who is likely to betray others, a police informant. Within cybercrime communities, however, we see an entirely different use of the term. Here, a rat is used to describe a commonly traded type of malware; a remote access trojan (or toolkit). This is why existing text mining and analysis tools may not perform well within such a domain-specific environment, with cybercrime argot being unique to the specific underground community.

Another term, 'leech', is used on the forum. Typically, this can refer either to the type of parasitic worm or to a person that sponges off others. Within technology, the second sense of the word is commonly used by people who carry out torrenting, to describe users that receive torrents without contributing back to the peer-to-peer network, used to shame the users. This term is also in use on the forum in a similar way, to refer to members that use the forum to solely ask questions without contributing back to other members in any way.

Terms used on the forum may also be entirely new. For example, 'ewhoring' is discussed on the forum, and is used to describe a type of fraud, in which stolen or shared sexualised images are used to trick victims into believing they have paid for a virtual sexual encounter [11].

Despite argot being an important signal within criminal communities, we believe we are the first to test whether it is an indicator of trustworthiness within the cybercrime underground. One reason that argot is an under-explored trust signal is it cannot easily be measured. Given the specialised nature of argot, it differs across communities, and is often implicit. We aim to detect argot and explore its use as a signal of trustworthiness. In this work, we explore the following research questions:

- How can natural language processing (NLP) techniques be used to efficiently detect argot usage on forums?
- Is there a relationship between the use of argot and reputation?
- How does the level of argot and reputation used by members vary over time?

2 Background & Related Work

2.1 Argot and the Criminal Underworld

The term argot is used to describe a specialised language. Originally the term was used to describe the particular ways in which criminals communicated, a useful way to prevent outsiders from understanding what is being said. The linguist Maurer provides a number of fascinating case studies in the 1930s and 1940s into the use of argot by particular types of law breakers and other shady figures. These include pickpockets [12], prostitutes [13], professional gamblers [14], narcotic addicts [15], moonshiners [16], forgers [17], and grifters [18] or con artists [19].

We will be using argot in its original context, that is, specialised language used within deviant subcommunities. However, the term has become more general over time, and is often used interchangeably with slang and jargon.

2.2 Reputation on Underground Forums

Communities on underground forums may use a reputation system as a proxy for trust. Dupont et al. [6] looks specifically at a system on a hacking forum, which uses a weighted approach for feedback. Members of a higher status have greater impact on a user's score: a new user posting positive feedback awards 1 point, whereas a moderator can award 5 or 10 points. They find that only a small fraction of forum members participate in the reputation system, and beginners are over two times more likely to report positive feedback of members compared to administrators.

Reputation systems can help members to establish trust on forums. Yip et al. [5] explored trust among cybercriminals on carding forums, finding one key challenge of needing to determine if another forum member can be a trusted individual, a dishonest trader ('ripper', who may provide worthless goods or sell products with backdoors [6]), or a law enforcement associate.

To combat 'rippers' on forums, Dupont et al. [6] note a sanction system used on the forum. Administrators may completely remove all of a user's positive reputation feedback on the forum, leaving only negative reputation feedback. However, Lusthaus [8] finds that such sanctions are not as useful, since there is no longer a large cost to switching profile: if a member has negative reputation, they can lose this negative signal by creating a new forum profile. Lusthaus compares this to conventional crime, where individuals have a known identity and need to increase anonymity, whereas in cybercrime, individuals start with no identity.

Work by Holt et al. [7] looked into the role of trust signals in cybercrime marketplaces for stolen data. They use a zero-inflated Poisson regression model to explore the relationship between marketplace signals and reputation received. One finding showed that having negative feedback correlates with receiving more positive reputation votes. They hypothesise that this can be due to either a seller having a large enough user base,

or rippers using positive feedback to obscure negative feedback.

Reputation on forums has also been used as a validation metric [20], as a proxy for trust. However, little work has explored the limitations of this metric, as forum members may game this system to appear more trustworthy than they are.

2.3 The Cold Start Problem

Established members on a forum will have had time to build a reputation and gain trust among other members. New members will start with a blank profile, and have to gain trust and reputation. However, since new members start with zero reputation, it is non-trivial to gain trust. We refer to this problem as the cold start problem.

The only prior work into overcoming the cold start problem within the cybercrime underground is by Vu et al. [21], in relation to a cybercrime marketplace. They use a combination of clustering and regression, to identify the group of members who overcome the cold start problem, and find that the majority of members build their reputation by participating in low value exchange-type transactions for exchanging currency.

2.4 Argot Detection in Cybercrime Communities

Argot detection typically uses a comparison between a base and target corpus, to identify words which may be out-of-dictionary or are used in a different context.

One approach is proposed by Seyler et al. [22], who aim to map 'dark' jargon to 'clean' jargon to make sense of new slang terms used on cybercrime forums. They compare two methods: KL-divergence and cross-context lexical analysis (CCLA), finding that KL-divergence outperforms CCLA on their simulated dataset. They also use their approach on a real-world corpus, however as they are unable to identify false negatives with their approach, they only validate the output of the top 'dark' jargon words. The authors use forum and Reddit datasets, where Reddit is used as a source of 'clean' jargon for a control. While they build word vectors for both the forum and Reddit datasets, our approach uses pre-trained word vectors for the 'clean' sample, reducing the time needed to collect and train on non-underground forum datasets. We use this '**DarkJargon**' approach as our baseline.

A different approach is taken by Yuan et al. [23], who aim to identify 'obfuscated' words. They use an approach based on word2vec [24], however they find that comparing word vectors from two separate models does not work. Instead, they propose a change to word2vec's skipgram model, which concatenates the word vectors from one hot encoding. This is equivalent to changing the dictionaries of the two corpora, prepending "A" to words used in one corpora and "B" to words used in another corpora. For evaluation, they find the approach has a precision of 0.91, but recall is 0.772. Therefore, of the predicted words, these are likely to be correct, but the approach was not successful in predicting positive words. In addition, the evaluation was limited to just drug and cybercrime product names, and requires collecting a large 'clean' corpus for model training.

Our method for argot detection utilises word embeddings. These place words into a space, providing semantic representations learned from examples of usage. While we use Euclidean space for our embeddings, other approaches have used different types of spaces for the task of hypernym detection [25]. However, when word embeddings are created for two separate corpora (e.g. cybercrime forum and Reddit posts), comparisons between them are not meaningful. Marchisio et al. [26] address this problem by comparing Euclidean and graph-based alignment methods, for transforming the word embedding spaces. They find that their performance varies on context. In our work, based on their results, we use the Euclidean approach to align the two embedding spaces, using a set of annotations.

Some approaches for argot detection use supervised machine learning models. While these have acceptable performance for test cases, Querioz et al. [27] highlights the issue that if models are used over longer periods of time, performance can degrade as lexical changes are introduced to the forum. The authors suggest a relabelling approach could be used, by labelling new data.

Further methods could be explored to support evolving lexicons. Ryskina et al. [28] analysed how new words are likely to be formed, finding that both semantic sparsity (surrounded by few words in the embedding space) and the frequency of growth rate both are predictive of this. Hamilton et al. [29] aligned vector spaces across the corpora representing different time periods, to measure which words have changed in meaning over time.

3 Ethics

Ethics approval was granted from the department's ethics committee for this work. We used data collected from a publicly available forum, and could not gain informed consent from all members as this would be considered to be spamming. As we only analyse posts and reputation data as a collective whole, rather than identifying individual users, under the British Society of Criminology's Statement of Ethics, this falls outside of the requirement of informed consent. We also avoid publishing details that could identify individuals, including usernames and original post contents.

4 Method for Argot Detection

The following section describes the dataset, the annotation process, the baseline comparison method, and the method used for argot detection. An overview of this is shown in Figure 1.

4.1 Data

We use a subset of the CrimeBB dataset [30], available for researcher use from the Cambridge Cybercrime

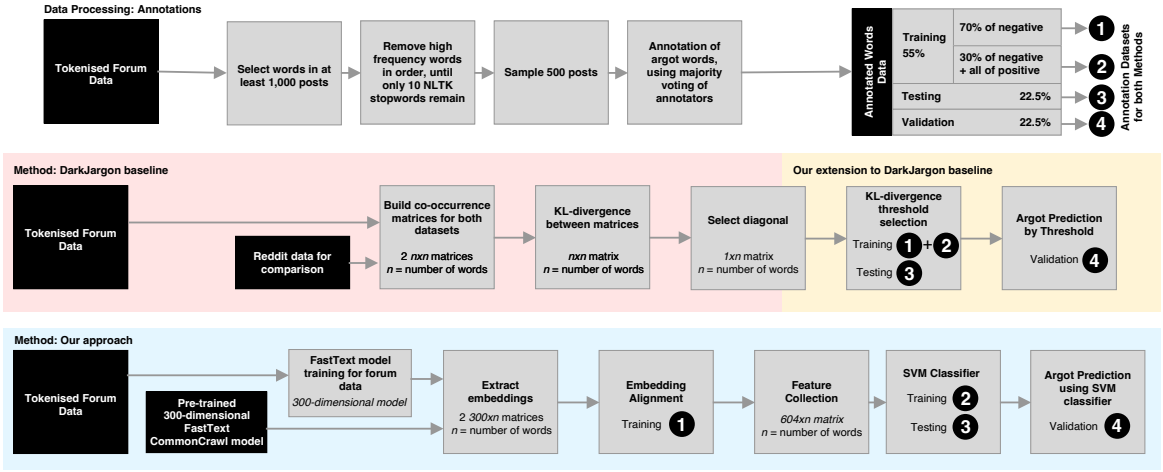

Figure 1: Comparison of our approach to the baseline method. Data from the annotation stage is used in different parts of methods for the DarkJargon baseline and our approach: training sets ❶ and ❷, testing set ❸, and validation set ❹. More details of the split are in §4.4

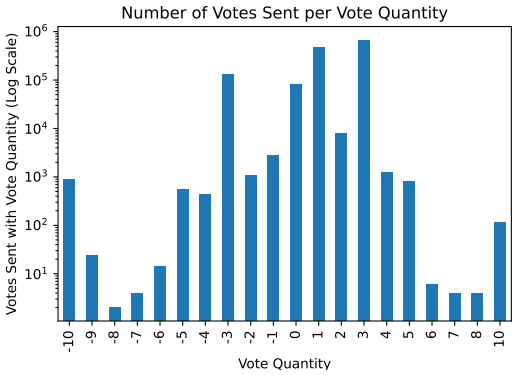

Figure 2: Number of Votes Sent per Vote Quantity

Centre[1]. CrimeBB contains posts scraped from 27 underground and dark web forums related to cybercrime, with over 13 years of post data. Each forum is structured by subforums, based on general topics e.g. hacking methods or marketplace advertisements, and are set by the forum administrators. Each subforum contains threads, which are an ordered collection of posts focusing on a defined topic set by the first post in the thread, such as a particular tutorial the author is sharing. Later posts can reply to the original first post, a reply to a later post by another user, or new information on the topic. While threads are typically focused on a particular topic, longer threads may become off-topic.

We use the HackForums subset of the database, an underground hacking forum on the surface web discussing various aspects of hacking techniques. Our dataset contains data since 2007 with over 190 administrator–curated subforums, with 4 million threads, and 42 million posts, created by over 630,000 members of the forum. Of these members, over 13,000 members contributed with at least six years of activity.

We use a dataset of reputation votes from HackForums in this work. Reputation votes are either positive or negative values between -10 and +10 sent from a user and 'received' by another (however, the receiving user does not have a choice to reject these). Figure 2 shows the number of each quantity amount sent on the forum. The value amount a user can send depends on their ranking on the forum [6]. We manually analysed the reputation votes sent, and found this changed in 2017, as scores were reset by the forum administrators due to misuse. Misuse included members sending negative reputation even though they had not traded with them, and members sending positive reputation to their friends regardless of their actions. Following this reset, members used automated scripts to quickly send other members reputation votes to try to recreate pre-reset scores. Therefore, we only use reputation data prior to 2017 in our analysis. Following this, the forum introduced a new contract system [21] as a new mechanism for trust. Contracts enable members to have a list of transactions that have taken place, for other members to observe, however there is no guarantee that these transactions reflect real-world transactions. We run a second analysis against confirmed received contracts (contracts which both parties have marked as being fulfilled), between the period 2018-06-11 and 2020-06-11. The contract system provides a transparent log of exchanges between users, and during this time, the reputation system was no longer in use.

While reputation scores can be problematic, they are the strongest signal as to trustworthiness available on cybercrime forums. We have omitted the period after 2017, where there were varying volumes in reputation, and spikes in activity, indicating reputation gaming. Before this, the volume of trust reputation voting was relatively stable. Trust is trust within members, and reputation is the main explicit method by which users vouch for others.

4.2 Tokenisation

We extract the contents of each post, and remove blocks including *URLs* and *images*. Then, we use NLTK's [31]

[1] `https://www.cambridgecybercrime.uk`

TweetTokeniser to tokenise the text, and join tokens (a meaningful group of characters, i.e. words) with whitespace. This is saved as a new column in a database table. This method enables the use of `ts_vector` in SQL statements using 'simple' (whitespace) mode to extract and summarise tokens. If Elasticsearch is instead used, a query can be run to obtain tokens from the text.

4.3 Annotations

We construct a list of training words based upon usage in HackForums. We first select words in at least 1,000 posts, to remove low-frequency words likely to be misspellings. We then use NLTK's list of English stopwords (e.g., "the", "and", "in"), removing the highest frequency words in order until only 10 stopwords remain in the set. We then sample 500 words for annotation by subject matter experts, to indicate if a given word is argot. We chose 500 words to provide a suitable sample size for our task, which could be annotated by our annotators in a reasonable timeframe. Note for other alignment papers there is often an existing resource such as a dictionary. But in our case there is no such dictionary, meaning that we have the overhead of annotations, and hence our seed vocabulary is smaller than related work.

We used Fleiss' Kappa for an indication of inter-rater reliability, and our annotations scored 0.59, which has moderate agreement [32]. The task of labelling argot is inherently difficult, with differing annotations between annotators. Therefore, we use majority voting among the three annotators to create our annotated training set.

4.4 Annotation Split

Annotations are split into two training sets ❶ ❷ with 55% of words, one testing set ❸ with 22.5% of words, and one validation set ❹ with 22.5% of words, to ensure there are enough samples for each part of the pipeline. We further split the training set into two training subsets: the first ❶ contains 70% of the negative (non-argot) annotations for alignment of word vector spaces. The second training subset ❷ contains the remaining 30% of negative annotations combined with the positive argot annotations, for training the support vector machine (SVM) classifier.

For the baseline approach, we combine the two training subsets ❶ + ❷ into a single training set, as there is only one training step.

4.5 Baseline Comparison

We compare our approach to DarkJargon [22]. The authors use two methods to identify hypernyms (more general related words) of slang: KL-divergence and Cross-context Lexical Analysis (CCLA), finding that KL-divergence outperforms CCLA. We use the KL-divergence method to build a baseline to compare our approach to. KL-divergence is used to measure the divergence of word co-occurrence between the HF corpus and baseline Reddit corpus, to provide a proxy for measuring context words are used in.

First, we build co-occurrence matrices for HackForums and Reddit data from Pushshift [33] (sampling data from January 2020, due to the large size of the Pushshift dataset). We use a window size of 10 (21 items in a window), as this allows us to compare results to the DarkJargon approach. Note that we limit both co-occurrence matrices to use a dictionary of HF words only, to focus only on HackForums data, and reduce the overall size of the matrices. We use Laplace smoothing, selecting an alpha of 0.1 to maximise the mean reciprocal rank (MRR) score, which is used to evaluate the order of results.

The authors' primary goal is to identify relevant hypernyms (general related words), rather than argot. To extend the task to identify argot, we add an additional step to the method: we place a threshold on the KL-divergence metric to select words used in different contexts between HackForums and Reddit.

To achieve this, we use the combined training set ❶ + ❷ and test set ❸ for tuning. We take the diagonal of the KL-divergence matrix from DarkJargon, to get the KL-divergence metric between the same word across the two corpora only. We sort the KL-divergence of HackForums words in descending order, to incrementally lower the threshold of KL-divergence for predicting argot. For each increment, we calculate the F1 score. We use this approach as it is more likely argot will be used in different contexts in HackForums compared to Reddit, due to the varying conversation topics and jargon across different platforms. Therefore, initial steps of decreasing threshold with increase F1 score, up to a given point, and then decrease as more common words are within the lowered threshold. We select the threshold which has the greatest F1 score: KL-divergence ≥ 0.91517.

Using our validation dataset ❹ with the DarkJargon approach, the accuracy score is 0.670 and the F1 score is 0.654.

4.6 Our Approach

4.6.1 FastText models

We use two FastText [34, 35] models for creating embeddings. The first is a model pre-trained on CommonCrawl data, which we use for comparison. For the second model, we train this over the dataset of tokenised posts. Parameters used are: vector size=300, window=5, min count=100, continuous bag of words (CBOW), epochs=5. These are selected as they have the same parameters for the pre-trained CommonCrawl FastText model [36].

We use a FastText model as this enables us to train the model on a large corpus with low computational resources. Alternatively, modern NLP approaches such as BERT [37] could be used instead, provided later annotations of argot are for tokens in context. In addition, BERT embeddings would need to be generated for every word in context for prediction, which would be computationally expensive (requiring GPU) compared to our

lightweight approach.

4.6.2 Create embeddings

Using the FastText models trained on CommonCrawl and HackForums data, we obtain word embeddings for all tokens in the HackForums dataset which have over 100 occurrences across all posts (including multiple usage in a post). We threshold tokens as computing statistics for all tokens can be computationally expensive, and those with low frequencies could be spelling mistakes of common words.

4.6.3 Align embeddings

We align the word vectors using Procrustes method, with the training data ❶ of only negative (non-argot) tokens set aside for alignment of the word vector spaces.

4.6.4 Feature collection

We also obtain definitions for tokens from the Urban Dictionary API[2], to create features for the number of definitions a word has, and the number of votes the top definition has.

The features we use are: cosine similarity of aligned vectors, distance of aligned vectors, HF word vectors (aligned), CommonCrawl word vectors (aligned), number of votes for the top definition in Urban Dictionary (or zero if not present), and number of definitions in Urban Dictionary (or zero if not present).

4.6.5 Predicting argot

We calculate the cosine similarity of words across the word vector spaces. We train a Gradient Boosting classifier [38] using features of the second training set ❷ and test set ❸.

4.7 Comparison

On the validation set ❹, our method has an **accuracy score of 0.723** and an **F1 score of 0.703**. This outperforms the DarkJargon baseline with an accuracy score of 0.670 and and F1 score of 0.654. Note that the detection of argot is a non-trivial task, requiring contextual information and domain knowledge to correctly identify words, including where a word may have multiple senses (usage).

5 Measuring Argot and Reputation

First, we test the hypothesis that there is a relationship between the use of argot and reputation votes. We then conduct further exploratory analyses to understand this relationship in greater detail, exploring how it varies over time.

[2]https://api.urbandictionary.com/

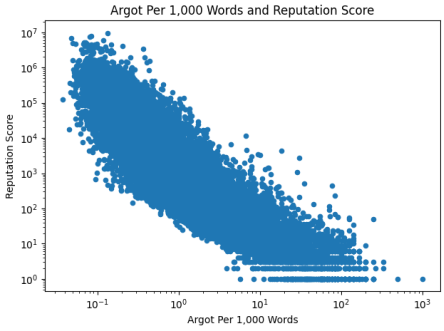

Figure 3: Argot Count Per 1,000 Words and Reputation

5.1 Relationship Between Argot and Reputation

We create an argot score for each user. We use the approach outlined in §4 to create a list of argot. Using this list, we create a materialised view in the database. We summarise the count of argot words used for each member.

When testing the hypothesis that there is a relationship between the use of argot and reputation votes, we control for the overall number of words posted by a user. We use this control variable as the more words a user posts, the greater the amount of argot they are likely to use, independent of their reputation.

Reputation votes are votes sent between a pair of users, with a given sender and recipient, and a positive or negative value. These are reputation votes received prior to 2017 (§4.1). We sum the quantity of votes to create a variable of reputation score.

5.1.1 Analysis with Reputation

We use the following variables for analysis (per user): number of argot words used, number of words posted by a user, and reputation score.

We first test if there is a correlation between argot words per 1,000 words used and reputation score. This uses cross-sectional data variables. We take the sum of reputation votes received prior to 2017-01-01, the count of argot used prior to 2017-01-01, and the number of words posted prior to 2017-01-01. We control for the number of words through dividing the count of argot words used by the number of words posted, multiplied by 1,000, to create a variable for the average number of argot words used per 1,000 words. We use Kendall's tau-b correlation to carry out the statistical test, as it is a non parametric measure of rank correlation, and can support ties in the data which Spearman's rank is unable to do.

Kendall's tau-b correlation was computed to assess the relationship between the reputation score and the average argot per post, for members (n=29,141) that have a reputation score and average argot per post $>$ 0, shown in Figure 3. There was a significant strong negative correlation between the two variables, $\tau_b = 0.740$, $p < 0.001$.

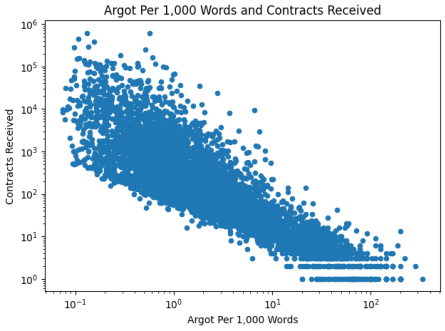

Figure 4: Argot Count Per 1,000 Words and Contracts Received

5.1.2 Analysis with Contracts

Second, we test if there is a correlation between argot words used per 1,000 words and contracts received. Contracts data is used instead of reputation data for this period, as the reputation system is no longer used. We take the sum of contracts received, the count of argot used, and the number of words posted during the period 2018-06-11 to 2020-06-01. This period is selected as the start of the contracts system (reputation system has been reset, more information in §4.1).

We assess the relationship between the count of contracts received by members, and their average argot per 1,000 words during this period, shown in Figure 4. Only members that have both a positive number of contracts and average argot per 1,000 words were included (n=6,159). There was a significant strong negative correlation between the two variables, $\tau_b = 0.699$, $p < 0.001$.

5.2 Argot and Reputation Over Time

In this section, we explore the relationship over time between argot and reputation. We use clustering for analysis. Initially, we used Group-Based Trajectory Modelling (GBTM) [39], however the proc traj library does not handle large datasets and we could not get the package's statistical models to fit to the dataset. We then explored the use of k-means longitudinal (R library kml) [40]. This uses the Lloyd's algorithm in k-means, which is run multiple times to avoid finding a local optimum, and additionally can fill in missing data values. We also explored the use of scikit-learn's [38] k-means implementation in Python, as we do not have missing values in the dataset. However, k-means does not work well with outliers, and expects variables to be normalised, which would become an issue when working with multi-trajectory clustering.

Therefore, we use Gaussian Mixture Models [38] to identify trajectory groups in the variables. Gaussian Mixture Models can be a more general version of k-means, and are equivalent to using GBTM with a normal distribution. Note that the second stage of GBTM – fitting a polynomial curve to trajectories – is not included here, as (1) we found polynomials did not fit the trajectory means well, and (2) our dataset uses data for consecutive months, and therefore we do not need to estimate values for missing months.

There is a large variation of number of months active for each member, during their first two years. Given the variation across this data, with a large number of members active for only one month and a small number of members active for several months, we select four groups: members with 1 (inclusive) to 6 (exclusive) months of activity (n=469,714), members with 6 (inclusive) to 12 (exclusive) months (n=37,003), members with 12 (inclusive) to 24 (exclusive) (n=24,541), and members active for every month of their first two years (24 months activity) (n=2,092).

Within these groups, we run two additional steps to normalise the data. First, due to some members not being active for all of the months, we use linear interpolation with exponential weighted mean to fill missing data points. Second, we find that there is a difference of volume of activity within these groups, leading clustering algorithms to cluster on volume. However, we want to explore how members have changed over time within their own activity, not how this contrasts to other members. Therefore, we identify the month with the maximum Argot Per Post for each member, and then divide each month by this. This gives us features that are **proportions** for each member, rather than raw data.

We cluster these proportions within each activity level group, shown in Figure 5, using our Gaussian Mixture Model approach. Note that in each group, the number of months active is not always consecutive. For example, a member active in months 0 and 12 will belong in the 1 to 6 months group, and missing months 1-11 and 13 onwards will use linear interpolation to fill the inactive months. For members active between 1 and 6 months, the largest cluster contains a high continuous level of argot used by members. The second largest cluster has the greatest argot usage for the first six months, before reaching a steady level. The fourth largest cluster contains members who have posted no argot.

For members active between 6 and 12 months, the four clusters have smaller differences between them over time. The largest cluster gradually increases and decreases in use of argot during the first year, before continuing at a steady level.

Members active between 12 and 24 months have greater diversity in trends over time. The largest cluster also gradually increases and decreases in use of argot during the first year, before continuing at a steady level. The second largest cluster continually increases over time for the two years. The smallest cluster, containing 2,780 members, shows an interesting pattern where argot per post starts high, then continuously decreases over the two years.

The final group contains highly committed members: those who are active for every month of the two year period, which does not require interpolation to fill missing values. The largest cluster contains a consistent level of argot per post used by members over time. However, the middle cluster contains members with a decreasing level of argot over time.

5.3 Cold Start Problem

We next explore the cold start problem, in which new members try to overcome the issue of having zero reputation among existing members with established reputation. This can involve members engaging positively with other members, to gain reputation votes, or manipulation to quickly build reputation regardless of their activity. We explore the cold start problem for two groups (12 to 24 months and 24 months), as these both contain two declining clusters of Argot Per Post over time, to explore which group variables correlate with this pattern. The two groups with less than 12 months of posting do not contain a declining cluster, therefore we do not analyse these groups.

Mean variables for members active between 12 and 24 months are shown in Figure 6. For the smallest cluster, which shows a continual decrease of Argot Per Post over time, we find that the argot per post per month sharply decreases within the first five months, while the reputation score increases.

Mean variables for members active for 24 months is shown in Figure 7. There is a similar pattern to the previous group, in which the middle cluster contains members with a decreasing level of argot over time. We observe a significant decrease in argot per post at the same time as a significant increase in reputation score. In addition, this group is characterised by a greater number of posts per month and argot per month than other clusters.

Across these two declining groups, they both use a high level of argot in their forum postings, and this decreases once reputation metrics increase. These two groups have thus used argot as a way to overcome the cold start problem.

Note in these figures, we are not clustering or correlating between these variables, which could introduce multicollinearity issues. The purpose of these figures is to highlight the variables for each detected cluster, in order to explore why argot is decreasing in one group while rising in another.

6 Discussion & Future Work

First, we presented a method for detecting argot, which outperformed the baseline approach. While this approach improved the accuracy and F1 scores on a strong baseline, future work in argot detection could focus on using more advanced NLP models to further improve metrics. For example, BERT could be used to identify argot. However, our annotations were per-word as we obtained a single word vector per-word, whereas BERT would need to train on argot words in context. This would require considerable more annotation time, and further, fine-tuning a BERT model to our task requires powerful GPUs, which we did not have available. Finally, in our task, we were able to identify argot words to provide counts of these words per-user for analysis. With BERT or similar pre-trained language models, we would have to predict argot usage for each sample in the dataset, instead of a straightforward database query. This would add significant overhead to analysis.

We explored the relationship between both argot and reputation scores, and argot and received contracts, finding both have a strong negative correlation. By analysing trajectories of groups of users over time, we also found two groups of members who have decreasing argot usage while their reputation increases. This can indicate signalling between forum members, where argot is initially used as a signal of trust by displaying knowledge of words used within this community. Later, as reputation increases, this becomes the main trust signal, and usage of argot decreases, thereby overcoming the cold start problem. The cold start problem applies where members aim on increasing their reputation score, in order to improve trustworthiness for trading, before reducing argot usage once trust is established. There are also users in which they wish to continue using the community's lexicon in order to fit in, requiring an ongoing level of cognitive activity.

Future work in analysing the relationship between argot and reputation could explore if this pattern exists across other cybercrime forums, and use advanced modelling techniques to cluster multivariate datasets (e.g. changing argot level and reputation over time together). Also, later work could look at whether reputation or argot comes first – do members increase their argot usage as reputation increases, or use forum-specific argot to build their profile? This could analyse groups where members have increasing reputation first before using argot.

6.1 Limitations

We note that this work has a few limitations. Firstly, as the dataset uses scraped data based on real-world interactions, it will not be "perfect". We explored the reputation dataset before working on it, to identify major outliers in the dataset that could affect our result. This included the reset of the reputation system in 2017 (§4.1), leading to members immediately sending a significantly greater level of reputation votes to each other, in order to try to recreate their previous score. Changes such as these can affect overall analysis, and we therefore chose to exclude this period, and include a second period using data from received contracts.

Also, we note that our approach uses a bag-of-words approach. We identify a list of argot, to measure usage. However, these argot words do not contain contextual information. Future work should use more advanced NLP models which take context into account, to improve prediction accuracy and measurements.

7 Conclusion

In this work, we presented a method for efficiently detecting argot usage on forums, carried out a cross-correlation analysis between argot and reputation, and explored the cold start problem with the reputation system. Our argot detection method combines pre-trained word embeddings with forum-specific embeddings, using an alignment approach with a set of annotations.

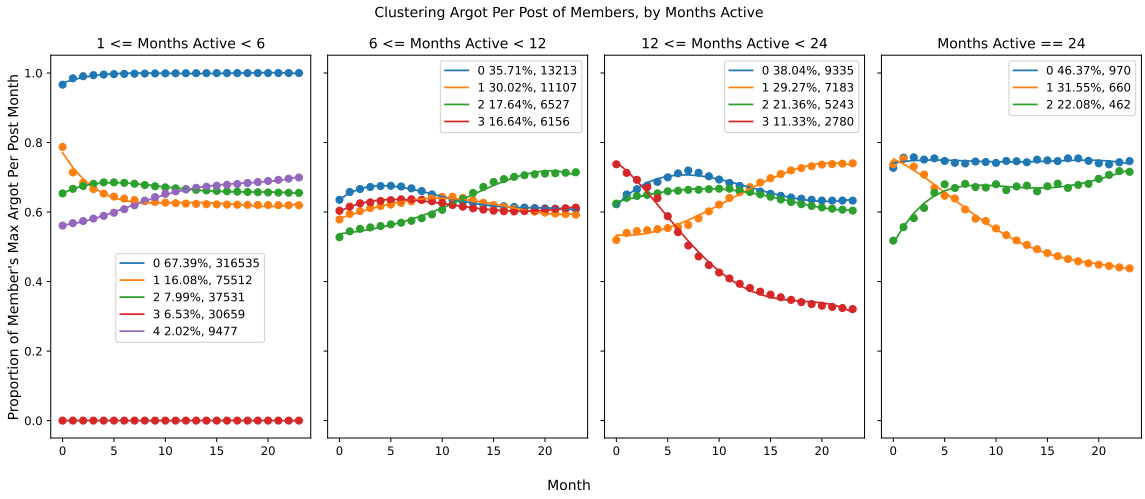

Figure 5: Clusters of Argot Per Post of Members, Over Different Activity Levels

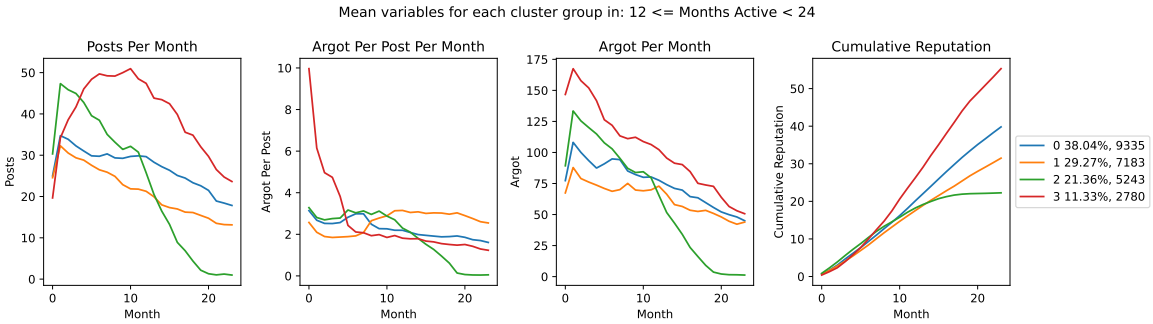

Figure 6: Mean of Variables for Members Active between 12–23 Months

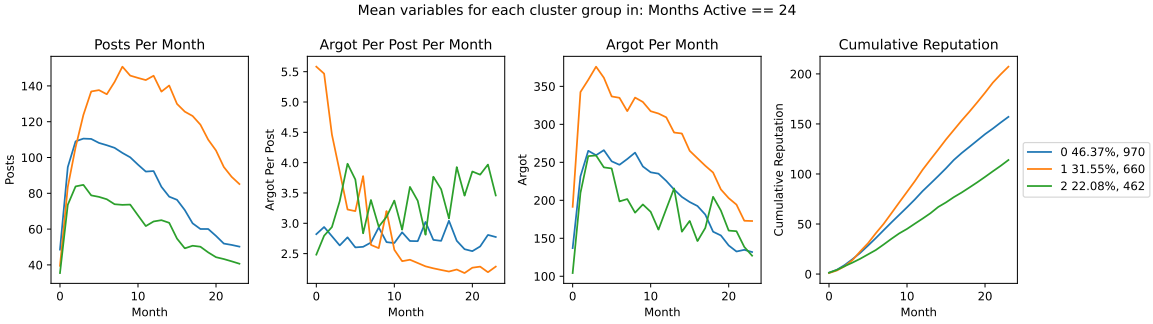

Figure 7: Mean of Variables for Members Active for 24 Months

Argot usage can be obtained from the dataset using simple database queries, requiring little computational time. We use our method on a subset of CrimeBB to explore the usage of argot over time. We used Kendall's tau-b correlation with between argot per post and reputation votes, and between argot per post and received contracts, both finding a significant result. Finally, we used clustering to explore how argot and reputation usage varies over time among forum members, and how members overcome the cold start problem. We find that as time passes and reputation increases, argot usage decreases.

Acknowledgements

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949127)

References

[1] D. Gambetta, *Codes of the Underworld: How Criminals Communicate*. Princeton University Press, 2011.

[2] P. Baker, *Polari-the lost language of gay men*. Routledge, 2003.

[3] S. A. Bakken, "Drug dealers gone digital: Using signalling theory to analyse criminal online personas and trust," *Global Crime*, vol. 22, no. 1, pp. 51–73, 2021.

[4] J. Lusthaus, "Industry of anonymity," in *Industry of Anonymity*, Harvard University Press, 2018.

[5] M. Yip, C. Webber, and N. Shadbolt, "Trust among cybercriminals? carding forums, uncertainty and implications for policing," *Policing and Society*, vol. 23, no. 4, pp. 516–539, 2013.

[6] B. Dupont, A.-M. Côté, C. Savine, and D. Décary-Hétu, "The ecology of trust among hackers," *Global Crime*, vol. 17, no. 2, pp. 129–151, 2016.

[7] T. J. Holt, O. Smirnova, and A. Hutchings, "Examining signals of trust in criminal markets online," *Journal of Cybersecurity*, vol. 2, no. 2, pp. 137–145, 2016.

[8] J. Lusthaus, "Trust in the world of cybercrime," *Global Crime*, vol. 13, no. 2, pp. 71–94, 2012.

[9] C. Herley and D. Florêncio, "Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy," in *Economics of Information Security and Privacy*, pp. 33–53, Springer, 2010.

[10] J. Franklin, V. Paxon, A. Perrig, and S. Savage, "An inquiry into the nature and causes of the wealth of internet miscreants," in *Proceedings of the 14th ACM Conference on Computer and Communications Security*, CCS '07, (New York, NY, USA), p. 375–388, Association for Computing Machinery, 2007.

[11] A. Hutchings and S. Pastrana, "Understanding ewhoring," in *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 201–214, 2019.

[12] D. W. Maurer, *Whiz mob: A correlation of the technical argot of pickpockets with their behavior pattern*. Rowman & Littlefield, 1964.

[13] D. W. Maurer, "Prostitutes and criminal argots," *American journal of sociology*, vol. 44, no. 4, pp. 546–550, 1939.

[14] D. W. Maurer, "The argot of the dice gambler," *The ANNALS of the American Academy of Political and Social Science*, vol. 269, no. 1, pp. 114–133, 1950.

[15] D. W. Maurer, "The argot of the underworld narcotic addict," *American Speech*, vol. 11, no. 2, pp. 116–127, 1936.

[16] D. W. Maurer, "The argot of the moonshiner," *American Speech*, vol. 24, no. 1, pp. 3–13, 1949.

[17] D. W. Maurer, "The argot of forgery," *American Speech*, vol. 16, no. 4, pp. 243–250, 1941.

[18] D. W. Maurer, "The argot of the three-shell game," *American Speech*, vol. 22, no. 3, pp. 161–170, 1947.

[19] D. W. Maurer, "The argot of confidence men," *American Speech*, vol. 15, no. 2, pp. 113–123, 1940.

[20] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, "Characterizing Eve: Analysing cybercrime actors in a large underground forum," in *International Symposium on Research in Attacks, Intrusions, and Defenses*, pp. 207–227, Springer, 2018.

[21] A. V. Vu, J. Hughes, I. Pete, B. Collier, Y. T. Chua, I. Shumailov, and A. Hutchings, "Turning up the dial: the evolution of a cybercrime market through set-up, stable, and COVID-19 eras," in *Proceedings of the ACM Internet Measurement Conference*, pp. 551–566, 2020.

[22] D. Seyler, W. Liu, X. Wang, and C. Zhai, "Towards dark jargon interpretation in underground forums," in *Advances in Information Retrieval* (D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, eds.), (Cham), pp. 393–400, Springer International Publishing, 2021.

[23] K. Yuan, H. Lu, X. Liao, and X. Wang, "Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces," in *27th USENIX Security Symposium (USENIX Security 18)*, (Baltimore, MD), pp. 1027–1041, USENIX Association, Aug. 2018.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Advances in Neural Information Processing Systems 26 (NeurIPS)*, 2013.

[25] M. Le, S. Roller, L. Papaxanthos, D. Kiela, and M. Nickel, "Inferring concept hierarchies from text corpora via hyperbolic embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3231–3241, Association for Computational Linguistics, July 2019.

[26] K. Marchisio, Y. Park, A. Saad-Eldin, A. Alyakin, K. Duh, C. Priebe, and P. Koehn, "An analysis of Euclidean vs. graph-based framing for bilingual lexicon induction from word embedding spaces," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, (Punta Cana, Dominican Republic), pp. 738–749, Association for Computational Linguistics, Nov. 2021.

[27] A. L. Queiroz, B. Keegan, and S. Mckeever, "Moving targets: Addressing concept drift in supervised models for hacker communication detection," in *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pp. 1–7, 2020.

[28] M. Ryskina, E. Rabinovich, T. Berg-Kirkpatrick, D. Mortensen, and Y. Tsvetkov, "Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods," in *Proceedings of the Society for Computation in Linguistics 2020*, (New York, New York), pp. 367–376, Association for Computational Linguistics, Jan. 2020.

[29] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 1489–1501, Association for Computational Linguistics, Aug. 2016.

[30] S. Pastrana, D. R. Thomas, A. Hutchings, and R. Clayton, "CrimeBB: Enabling cybercrime research on underground forums at scale," in *Proceedings of the 2018 World Wide Web Conference*, pp. 1845–1854, 2018.

[31] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st ed., 2009.

[32] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[33] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 830–839, May 2020.

[34] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[35] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[36] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," 2017.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] D. S. Nagin and C. L. Odgers, "Group-based trajectory modeling (nearly) two decades later," *Journal of Quantitative Criminology*, vol. 26, no. 4, pp. 445–453, 2010.

[40] C. Genolini and B. Falissard, "Kml: k-means for longitudinal data," *Computational Statistics*, vol. 25, no. 2, pp. 317–328, 2010.