University of Vermont
ScholarWorks @ UVM

UVM Honors College Senior Theses | Undergraduate Theses

2019

"It's So Bomb": Exploring Corpus-Based Threat Detection on Twitter with Discourse Analysis

Addie Beach
University of Vermont

Follow this and additional works at: https://scholarworks.uvm.edu/hcoltheses

Recommended Citation
Beach, Addie, ""It's So Bomb": Exploring Corpus-Based Threat Detection on Twitter with Discourse Analysis" (2019). *UVM Honors College Senior Theses*. 298.
https://scholarworks.uvm.edu/hcoltheses/298

This Honors College Thesis is brought to you for free and open access by the Undergraduate Theses at ScholarWorks @ UVM. It has been accepted for inclusion in UVM Honors College Senior Theses by an authorized administrator of ScholarWorks @ UVM. For more information, please contact donna.omalley@uvm.edu.

“It’s So Bomb”: Exploring Corpus-Based Threat Detection on Twitter with Discourse Analysis

Addie Beach
Department of Romance Languages and Linguistics
Honors College Thesis
University of Vermont
Spring 2019

1. Abstract

As social media increases in popularity, its ability to create culturally meaningful tools grows as well. One of the most promising tools is categorization software, which analyzes the linguistic data in social media posts to make predictions. It does with the help of corpus linguistics, a form of analysis designed to pick out the most frequent and/or significant words in a dataset. This study focuses on software intended to detect threats. While this technology has the potential to flag threatening language used by groups or individuals, the text search strategies it currently uses often result in a high number of false positives, making it too unreliable for effective use. The software is most effective at marking whether or not a specific word is present in a tweet, not determining whether or not this word is actually being used in a threatening way (e.g. "I'm planning on killing him" vs. "this silence is killing me"). Discourse analysis, which looks at the role context plays in language, could minimize these errors by helping researchers refine the software in a manner that more closely matches how people actually use language. The goal of this project, then, is to investigate ways of combining corpus linguistics and discourse analysis with a Twitter database to improve threat analysis.

2. Introduction and Literature Review

2.1. *Discourse Analysis and Corpus Linguistics*

For many linguists, the methodologies of discourse analysis and corpus linguistics have seemed incompatible. Discourse analysis necessarily relies on small-scale analysis, often focusing on the pragmatic conventions and underlying ideologies of a single conversation. Context is therefore crucial to discourse analysis, and this context is often lost in the broad reach of corpus analyses. As Virtanen (2009) points out, corpus linguistics emphasizes token frequency rather than token relevance. This shift in focus tends to reflect corpus linguistics' increased concern over representativity – the fact that discourse analysis has a sample of only a few speakers decreases its generalizability (Virtanen, 2009). While discourse analysis treats the speech of individuals as key to understanding linguistic trends, corpus linguistics tends to locate these insights at a broader population level. Ultimately, these differences reflect what appears to be a fundamental, irreconcilable rift between the two methodologies.

Yet, despite these differences, more and more researchers are starting to find ways in which the two approaches can support each other. Collocations seem to be one area where context and corpora often meet. Overall, collocations show how frequently two words co-occur in a dataset and can provide information about the relative rarity or popularity of a certain combination – for example, "of the" occurring more frequently in everyday speech than "of twenty." Studies have shown that collocations may be able to help us answer questions about how native speakers come to understand the meaning of a word based on its immediate environment, and how these judgements accordingly influence linguistic choices made in specific situations (Virtanen, 2009). Therefore, collocations allow us to view the implicit contextual links between words that contribute to sociolinguistic phenomena like style and register. Corpus linguistics allows us to locate larger linguistic trends, such as inter-genre variation, whose use and significance we can better understand through discourse analysis.

Researchers are already starting to find meaningful results from this relationship between the two fields. This is especially true of corpus linguistics and critical discourse analysis (CDA), which attempts to uncover underlying worldviews or perspectives present in discourse. Collocates can reveal common formulaic metaphors, topics that tend to cluster together, or positive or negative trends in lexical environments (called semantic prosody). A 2011 study looking at competing Islamic ideologies in Saudi-Arabia, for example, examined collocates of the word “Wahhabi” in two books presenting different opinions of the ultra-conservative movement (Salama). The researchers found that “Wahhabi” appeared with more neutral or positive collocates such as “discussion” and “vision” in the supporting text. At the same time, it was linked to more negative words, including “infiltration” and “cult,” in the opposing text, implying a more sinister view of the movement and indicating the stance of the writer (Salama 2011). The worldview and motivations of the author, then, were expressed through the semantic prosody of the collocates.

Furthermore, corpus-based CDA analyses can have proven themselves useful even when the position of the writer or speaker is not straightforward. Similar studies, such as that by Bednarek & Caple (2014), have been conducted looking at bias at seemingly more “neutral” sources like newspapers, often finding underlying opinions expressed through word choice. These researchers also found, through collocate and keyword analyses, that articles often contained words strongly linked to traditional “news values” that are thought to determine the readability of a piece. These values include superlativeness and negativity, demonstrated through phrases like “Britain’s worst nuclear accident” (Bednarek & Caple, 2014). Therefore, collocates can be a powerful tool for sentiment analysis, whether or not explicit expression of opinion is an expected feature of the genre. Additionally, the latter finding suggests a link between the conventions of a register and corresponding linguistic decisions.

Salama (2011) and Bednarek & Caple (2014) both incorporate elements of sentiment analysis, which attempts to identify positive, negative, or neutral attitudes in a given text. Sentiment analysis often uses collocates to locate words that may contribute the expression of these value-based opinions (e.g. positive ones like “bright” and “good” versus negative ones like "unhappy" and "wrong"). Much of collocation-based sentiment analysis ties into metaphor studies, a particularly promising area of discourse-corpus research that investigates the linguistic structure of metaphors. This research demonstrates the potential of corpora to deal with multiple aspects of meaning, specifically differentiating between non-literal and literal definitions of a word and exploring these definitions' use. Again looking at newspapers, Krennmayr (2015) identified some of the most common metaphors used in reporting. These include personification, as in "[the Labour Party] hopes to transform the situation" (Krennmayr, 2015). Others have made use of corpus linguistics' ability to analyze grammatical features, particularly when used in conjunction with part-of-speech tagging, a system used to automatically annotate words in a corpus according to their grammatical category. For instance, studies on metaphor use have looked at how morphosyntactic variations, such as changes in grammatical category, plurality, or tense, can have consequences for the meaning of a metaphor; compare, as an example, "the rock on which society is built" to "their marriage has been on the rocks" (Semino, 2017). This research can also help us trace the development of metaphors over time, showing how the usage of a word can move from concrete to abstract and vice versa – "dull," for instance, first having started out describing a lack of intelligence and then a lack of physical sharpness (contrary to what we would otherwise expect) (Semino, 2017).

Finally, corpora also allow us to make much larger analyses of the pragmatic functions of grammatical constructions than previously available. As discussed previously, most traditional discourse analyses are relatively small scale, both by design and out of necessity, rendering it difficult to make generalizable claims about the social meaning of grammatical features. However, combining discourse analysis with corpus research has allowed to do so to a greater degree, particularly in regard to tense shifts. In a 2019 study, Frazier & Koo applied discourse methods to a corpus of TV and radio interviews, investigating how placement of a verb in a conversation (e.g. at the end or beginning of a turn) correlated with tense. In doing so, they discovered that use of the present-perfect tense was overwhelmingly associated with actions of position-taking and support in conversations (Frazier & Koo, 2019). For example, while presenting his opinion on contemporary film scores, an interviewer stated that "the music we expect to hear in the movie theater…*has changed* a lot over the years" at the start of his introduction, thus using tense to situate his position in a broader historical context (Frazier & Koo, 2019). Therefore, the combined use of discourse analysis and corpora presents us with the opportunity to view how pragmatic variation is enacted by a wide community, allowing us to see underlying trends that would otherwise be difficult to locate with traditional discourse methods alone and then support these findings with quantitative data.

2.2. *Social Media as Corpora*

Given that corpus linguistics derives much of its strength from being able to interpret data that represents a vast population of speakers and genres, it makes sense that many researchers are increasingly looking to social media as a potential source of corpora. Social-media-based corpora differ from traditional sets in ways both beneficial and detrimental. On one hand, easy access to a continuously updating data source allows for both a more accurate picture of speech as it is currently used and a way to measure diachronic change in greater detail. At the same time, though, some question the legitimacy and representativity of the data (Simpson, 2017). Online communication does not necessarily adhere to the conventions of spoken or written speech, preventing findings from being generalized to offline communication, and limited (or inaccurate) metadata, often due to users disabling location trackers or falsifying their age to skirt restrictions, can make pinpointing a specific community challenging.

As such, more and more researchers are beginning to evaluate the potential applications and tradeoffs of social media in corpus analysis, as well as the relevance of these findings to sociolinguistics as a whole. In a 2017 study, for example, Simpson investigated the use of social media corpora as a way to inform Supreme Court decisions. Most courts have traditionally relied on the use of dictionaries to resolve conflicts of meaning in legal wording; for example, whether to "carry" a weapon applies just to having a weapon on your person or can extend to being held in a vehicle (Simpson, 2017). Simpson discovered that because they contain more information about the relevance of competing definitions, corpora in general allow for greater accuracy by providing insight into the "ordinary meaning" of language, or how it is used and understood by speakers. With access to an even greater number of speakers, Simpson found that this is especially true of web-based datasets, noting that Twitter's volume of text alone is unmatched by most traditional corpora, including the Corpus of Contemporary American English (Simpson, 2017). Social media therefore presents a unique opportunity for corpus linguistics research, often able to provide a better sense of language as it is actually practiced and potentially refining the applicability of these findings in the process.

In addition to giving us insight into how people are using language, social media research also allows us to see what people are talking about. Studies have found corpus methodology is capable of detecting significant events being discussed in a dataset, which can help data be further sorted into useful categories. As the speed at which social media corpora are able to be updated grows faster, they also bring us to closer to being able to observe reactions to major events during, soon after, or even before they occur. Matheson (2018), for example, used discourse-corpus methodology to investigate responses on Twitter soon after a major crisis. Looking at a set of 420,000 tweets produced in the 18 days following the Christchurch earthquake in New Zealand, all of which were identified by searching for relevant hashtags with corpus software, Matheson studied how responses to the crisis were enacted through language on Twitter. He found that users saw Twitter alternatively as a public platform for expressing political opinions, a more pragmatic communication service for offering emotional and physical support, and a community for negotiating shared experience (Matheson, 2018). This study demonstrates not just the degree to which social media conversations can provide us with insight on traumatic events, but also that more in-depth corpus studies are often needed to understand how users interact with the platform differently. As Matheson notes, Twitter presents people with an opportunity to simultaneously interact personally with other individual users and more publicly with the community as a whole, blurring conversational boundaries. As a result, social media communication represents a wide variety of registers, and may include public declarations mixed with more private discussion topics. As such, though corpus linguistics is often needed to sort out relevant information in these expansive datasets, discourse analysis is then necessary to understand the conventions and structure of social media interactions. When applied to analysis of major crises, we see the varying ways users interact with Twitter as a resource, showing how much variation can occur even when discussing a single topic. As the gap narrows between when an event occurs and when it is detected in internet discussions, we are better able to observe the evolution of these topic-based trends over time.

Though Matheson (2018) shows us the applications of corpus-assisted discourse analysis on an event-level scale, the corpus-discourse approach is also able to comment on more individual interactions as well. This is particularly true of research concerning threatening language. Indeed, researchers have identified threatening language that is in itself a violent action, employed by one user to target and intimidate another. Cyberbullying is one area of particular focus, both because of its prevalence in the current social media landscape and because early detection by moderators presents itself as a relatively manageable solution (Van Hee et. al., 2018). A 2018 study by Van Hee et. al. found that researchers were able to detect signals of cyberbullying on social media by manually categorizing potentially threatening posts according to the role they played in online conversations. These categories included clear signs of bullying, such as threat/blackmail, defamation, harasser encouragement, as well as more ambiguous markers like insult, which could potentially harmless tweets like “hi bitches” (Van Hee et. al., 2018). With this annotation, they were then able to identify which parts users were playing in the cyberbullying process: bullies, victims, or bystanders (Van Hee et. al., 2018). Overall, by classifying the linguistic features (often shared words or phrases) that defined each category, the researchers were able to develop a system that was able to automatically detect signals of cyberbullying (Van Hee et. al., 2018).

A key component of identifying threatening language is cataloguing and categorizing the uses that appear in a dataset. Van Hee et. al. (2018) did this through the roles the users were enacting in the cyberbullying practice. In doing so, they demonstrated that not all violent language is created equal, and it may differ along lines of severity or function in the discourse. Corpus linguistics can help identify violent keywords in the data, allowing for threatening tweets to be filtered out and flagging key lexical features in automatic detection software. However, understanding the significance of violent language instances – that is, whether or not they are threatening and what kind of threat they pose – usually requires discourse analysis.

Corpus-discourse analysis was a key part of a 2017 study by Clark & Grieve looking at abusive language use online. When considering tweets containing threatening language, they discovered that it often was not sufficient to sort the tweets based on the presence of offensive words (such as profanity) on their own. To do so ignores that these features both are not always present in abusive tweets and can be used in ways that are not abusive, leading to false positives (Clark & Grieve 2017). In response to this, the researchers posed a method of tweet classification based on multidimensional analysis (MDA). MDA considers the communicative purpose of a tweet, especially how a user expresses stance and the responses they try to elicit from others (Clark & Grieve 2017). With a corpus comprised of 1,486 tweets that had been coded as racist or sexist, Clark & Grieve attempted to locate basic discursive categories the tweets could fall into with MDA.

Overall, they found that the majority of the tweets could be described in terms of three dimensions: interactive, antagonistic, and attitudinal (Clarke & Grieve, 2017). Each category was defined by a matrix of grammatical features, such as verb tense and pronoun type. The tweets then confirmed either the presence or absence of these dimensions based on the set of features they contained. The interactive category was defined by users attempting to respond directly to either the Twitter feed as whole (using hashtags) or by singling out specific users. The dimension contained frequent use of first and second person pronouns, which explicitly designated the users in the discourse and the intended audience, and question features such WH-words and initial DO (Clark & Grieve, 2017). By contrast, tweets that scored low on the interactive scale tended to use more existential "there," acting as statements of fact rather that interrogations (Clarke & Grieve, 2017). Similar trends between form and content, as well as between positive and negative presence, were found in the antagonistic and attitudinal categories.

This link between content and form could have important implications for attempts to refine corpus-based analyses, allowing us to apply content-based trends to structure-based ones and create better categorization techniques.

2.3. *Corpus-Based Information Extraction*

The foundation of online corpus-based analysis is information extraction, the process through which data is sorted into meaningful categories. Overall, information extraction can be separated into three basic steps: tokenization and lexical processing, syntactic processing, and domain analysis (Bock, 2012). The first of these steps, tokenization, involves combing through the data and noting the part of speech (POS) of the tokens present. The next two steps in information extraction, syntactic processing and domain analysis, then seek to refine these POS-based definitions. Syntactic processing looks at surrounding linguistic information, often syntax or collocations, in order to gain a better understanding of what definition is being used. Domain analysis then considers potential variations in the data that could result in skewed information - knowing that “USA,” “US,” “the United States of America,” and “America” all refer to the same country would be important for determining its frequency in the data, for example (Bock, 2012). Overall, all three steps contribute to the main goal of information extraction – retrieving a set of workable, definition-based categories for the words in a corpus.

Though all information extraction software relies on these three basic steps, programs differ in the amount of manual human input they require (Small & Medsker 2013). For more supervised approaches, researchers must first analyze the patterns present in the data themselves, accounting for all possible variations that could occur. They then feed these findings into the program, which applies the rules to new datasets. On the other side of the spectrum, unsupervised programs are able to automatically extract patterns from the corpus using machine-learning derived techniques such as neural networks, with minimal researcher direction. While fully unsupervised programs are often cited as the ultimate goal for information extraction software, they are still limited in their application, despite promising advances in natural language processing (Small & Medsker 2013). Furthermore, no matter the approach, manual linguistic analysis is always needed at some stage of the process in order to identify relevant patterns off which the programs can work. Therefore, even with increasingly unsupervised programs, evaluating strategies for conducting linguistic analysis on corpora is still a necessary part of information extraction research.

Nevertheless, information extraction presents itself as a robust form of text analysis with applications across mediums and genres. Studies have shown that it is useful in analyzing news articles, for example, capable of extracting temporal data about events being discussed (whether they took place in the past or the future) based on inflectional morphology (Faiz 2006). Where this type of analysis is especially useful, however, is in online corpora (Small & Medsker 2013). Much larger than traditional spoken or written ones, web-based corpora present researchers with the chance to investigate language use on a broader scale than ever. For instance, Brigham Young University's iWeb corpus consists of around 14 billion words, 25 times the size of the Corpus of Contemporary American English, their largest non-web corpus (Davies, 2018).

Furthermore, information extraction has proven especially powerful when combined with language modelling techniques, particularly word embedding. Word embeddings allow researchers to map linguistic features according to their semantic content, with words that are more similar in meaning closer together and those that are dissimilar further apart. This type of analysis allows us to more clearly see the relationships between words in a dataset, and therefore makes it easier to conduct sentiment analyses that can be crucial to threat detection (Rezaeinia, Rahmani, Ghodsi, & Veisi, 2018). That being said, this software still often needs to be trained on datasets that have already been manually sorted by researchers. Additionally, it still has difficulty telling different, context-based definitions of a word apart – for example, knowing when “beetle” refers to an insect versus a car (Rezaeinia et al., 2018).

Ideally, then, we would be able to apply corpus-based threat detection software to distinguish between non-threatening and threatening definitions at the syntactic processing and domain analysis stages, filtering out irrelevant tweets and only marking those which contain legitimately threatening definitions (Bock, 2012). However, this only works if we understand both the linguistic environments that distinguish the threatening definitions of a word and all possible variations of a word that a dataset might contain. There are also several challenges that information extraction programs are only just beginning to address. One is the influence of sarcasm on language use. Because computer text lacks the intonational cues usually necessarily for locating irony in verbal speech, sarcasm detection relies almost entirely on context, which again often gets lost in the large-scale focus of corpora analyses (Muresan, Gonzalez-Ibanez & Ghosh, 2015). Adding onto this are complications that arise from domain analysis. While accounting for multiple variations of a word is an essential step in all corpus analyses, innovations are often seen in web-based text at a higher rate than other written mediums. Indeed, because many of these innovations are based on phonological features that originate in dialectal differences, web-based variations may be more similar to spoken ones that written ones, making it more difficult and time-consuming to account for them in corpora (Eisenstein, 2015). Therefore, when we look to improving the linguistic analysis that forms the basis of information extraction software, we must consider the important roles that context and lexical variation play in computer-mediated communication.

3. Methodology

This study is based on a corpus of 4,497 tweets which had been singled out by information extraction software, designed by the threat-detection company Social Sentinel, for containing lemmas, or variations on the base word "bomb" (e.g. bomb, bombing, bombed, etc.) The set contains both threatening ("I'm gonna bomb the office") and non-threatening ("her haircut is the bomb[1]") uses of the word. Most of the initial analysis focused on determining which uses of "bomb" the set contained. The data arrived pre-sorted into threatening and non-threatening categories by computer-based categorization that had been checked by human researchers, as well as pre-tagged for POS. However, the coding was occasionally inaccurate, with the threatening category containing non-threatening uses ("I managed to bomb two exams"). Even more importantly, though, threatening and non-threatening alone were not detailed enough categories for a more in-depth analysis of descriptive use patterns and the linguistic environment. Therefore, the data had to be manually coded to determine the full array of definitions present.

In order to analyze the "bomb" tokens in the data, they first had to be filtered out from the other data. This was done by using the keyword in context (KWIC) tool on the corpus analysis software package LancsBox (Brezina, McEnery, & Wattam, 2015). In addition to providing all instances of a token in the data, KWIC analyses give a limited string of words or characters found to the left and right of the tokens, thus also providing some information as to

[1] All examples from the data have been slightly paraphrased to protect user privacy

the context and linguistic environment in which the words appear. A search was done for all variations of "bomb." Each occurrence of the token was then copied into a spreadsheet, including some of the limited contextual information provided by the KWIC.

From here, the definition categorization process was then broken down into three steps. These consisted of (1) confirming the POS of the word, (2) noting any preliminary patterns in the linguistic environment or contextual clues, and (3) marking any variations of the words seen in the data. For the most part, these steps roughly correspond to the three stages of information extraction noted previously (Bock, 2012). The exception to this pattern is step 2. Step 2 in traditional information extraction is mainly focused on linguistic environment, specifically syntactic structure. Though this was considered in the analysis, the version used here also incorporated elements of discourse analysis methodology, most notably through an increased focus on context. As such, in addition to considering the grammatical function of the token, the role of the tweet of in the discourse or conversation taking place was considered as well. Contextual elements taken into account included the register of the tweet (e.g. a news headline vs. informal gossip) and any outside events or information that seemed to be referenced in the conversation (e.g. a baseball game, comedy show, concert, YouTube channel, etc.).

Step 1 was performed for every token, mainly checking to ensure the tagging was accurate and making corrections when it was not. Steps 2 and 3 were initially used to note general trends in the data from which more specific definitions were created. Some of these definitions relied mostly on POS and syntactic structure, such as the adjective category. Others were mainly based on context, like uses of "bomb" in sports slang. Most were a mixture of both, relying on both linguistic structure and context to define their categories, an example being the "movie" category of the verb "bomb" (e.g. "Harrison Ford film bombs at the box office"). This resulted in a list of 18 different definitions, both threatening and non-threatening. As the tokens were checked for POS, they were then also tagged for which of these definitions they best fit.

From here, sub-corpora were created based on these definitions. Some of the smaller, similar sets (such as the sports categories) were combined to have enough data to analyze, but the majority corresponded directly to their original definitions. These sub-corpora were then loaded into the LancsBox GraphColl tool (a collection of collocation-based statistical tests) for analysis, which consisted of two main parts. The first involved determining the most significant words in each dataset, primarily using raw frequency counts and mutual information scores. Raw frequency is the total number of times a token appears in a dataset. Mutual information (MI) scores, on the other hand, measure the strength and significance of the relationship between two collocates. The formula for calculating mutual information scores is as follows, where A is the frequency of the original word, B is frequency of the collocate, AB is the frequency of their co-occurrence, and span is the number of words analyzed to the left and right of the original word:

$$MI = \log((AB * sizeCorpus) / (A * B * span)) / \log(2)$$

While MI scores are a powerful tool for assessing how much weight to give collocates, they also have the tendency to favor obscurity over significance in smaller datasets, which many of these sub-corpora were (Lindquist 2009). With this in mind, raw frequency was used to filter out these errors, with the top collocates from each set being those that had both the highest MI and raw frequency ratio.

The next step considered pure raw frequency statistics on their own. This was done to ensure that the analysis took into account not just which words were unique, but which linguistic patterns were most common. The highest frequency words in any given dataset are likely to be common function words, so analyzing the corpora on raw frequency alone would not have provided much information regarding words unique or significant to the set (hence the use of MI scores). These function words, however, can provide us with insights as to the morphosyntactic structure: what articles most frequently occurred with what forms of "bomb," whether certain pronouns or prepositions were more common with one definition than other, etc. Furthermore, LancsBox provides data on the position of collocates in a sentence – that is, whether the collocate tends to appear to the right or left of the keyword – allowing for more in-depth syntactic analysis. In other words, both the significance testing and the frequency analysis were done to investigate the relationship between content and structural trends in the tweets, with the significance testing leaning more towards the former and the frequency the latter. From here, the key sub-corpora collocates could then be compared against each other to determine any consistent similarities or differences.

4. Data

Overall, eighteen principle uses of "bomb" were identified in the dataset as part of Steps 2 and 3. All definitions, including a brief description of each category and an example from the data, are listed in the chart below:

Table 1. Use-based categories for "bomb," with examples from the data

No.	Definition	Description	Example
1	Literal Nouns	"Bomb" referring to actual bombs	"NYU student arrested, pipe bomb found in backpack"
2	Literal Verbs	"Bomb" referring to dropping actual bombs	"I'm gonna bomb my office"
3	Baseball	Lexical items involving "bomb" in baseball	"Edwards hit a total bomb"
4	Golf	Lexical items involving "bomb" in golf	"They say bomb & gouge isn't golf"
5	Football	Lexical items involving "bomb" in football	"Y'all got excited by Brady throwing a 70 yard bomb"

6	Misc. Sports	Lexical items involving “bomb” in other sports	“WWE slam with a Badicka Bomb”
7	“Time bomb”	The fixed phrase “time bomb”	“It has been a ticking time bomb and dangerous to use”
8	“Truth bomb”	The fixed phrase “truth bomb”	“A truth bomb about McDonalds and sexism”
9	Euphemism	“Bomb” used as part of a fixed, euphemistic phrase	“This f bomb lmao”
10	Compound nouns	Nouns comprised of [bomb] + [noun], forming a new word	“Our most relaxing bath bombs”
11	Names	Names that contain “bomb”	“Now playing Hear My Cry by Khrisko Bomb”
12	Misc. Metaphors	Metaphors that include the word “bomb”	“We hit the deck like a bomb was incoming”
13	Historical/Political Bombs	“Bomb” used in a historical or political context	“Served with the 440th bombing squadron, 81st bomb group”
14	Adjectives	“Bomb” used as an adjective, either pre- or post-verb	“I’m educated and so bomb at what I do”
15	Bomb Squad	The phrase “bomb squad”	“LAPD bomb squad is responding”
16	Bomb Threat/Scare	The phrase “bomb threat” or “bomb scare”	“A bomb threat was found today”
17	Failure	“Bomb” used as a verb meaning “to fail”	“On my way to bomb my political science final”
18	Bomb-Related Nouns	Nouns comprised of [bomb] + [noun], referring to actual bombs	“Sending hundreds of Syrians into bomb shelters”

As the goal of this initial analysis was to capture the diversity of uses in the dataset, there are a few categories that either overlapped or could be combined. For example, Bomb Threat and Bomb Squad are both technically Bomb-Related Nouns. Generally speaking, they meet the criteria of (1) being compound nouns consisting of [bomb] + [noun] and (2) relating to actual bombs (whether they are real or fake is irrelevant, as long as they are not hypothetical or otherwise metaphorical). However, because this dataset was determined by use, and the phrases “bomb threat” and “bomb squad” constituted significant portions of the dataset by themselves, they were given their own categories in order to ensure that the makeup of the set was as reflected as accurately as possible.

A drawback of this approach, though, is that the categories were often too small to analyze. They rarely had enough tokens to justify basic corpus procedures, such as a collocation analysis. Therefore, these categories were further condensed into ten thematic categories: Sports, Names, Adjectives, Bomb-Related Nouns, Metaphors, Failure, Literal Nouns, Literal Verbs, and Compound Nouns. Most of them roughly correspond to their initial sets – the failure category for example, is comprised of the set of verbs meaning "to fail." The chart below displays the total definitions contained in each category:

Table 2. Thematic sets of use-based categories with proportion of total data

Category	Definitions	Bomb Tokens
Sports	Baseball, Golf, Football, Misc. Sports	539 (19%)
Names	Names	184 (6%)
Adjectives	Adjectives	772 (27%)
Bomb-Related Nouns	Bomb-Related Nouns, Bomb Threat, Bomb Squad	463 (16%)
Metaphors	"Time Bomb," "Truth Bomb," Euphemisms, Misc. Metaphors	175 (6%)
Failure	Failure	103 (5%)
Literal Nouns	Literal Nouns	53 (2%)
Literal Verbs	Literal Verbs	202 (7%)
Compound Nouns	Compound Nouns	114 (4%)
Historical/Political	Historical/Political Bombs	113 (4%)

The largest group in this list was the adjective category, with 13,859 total tokens (including non-bomb tokens), while the smallest was the explicit metaphor category at 2,221 (with an average category size of 5,728). Though these categories were still rather small for a corpus analysis, they were at least large enough to contain some diversity in form.

From here, the collocation analysis was performed to determine patterns present across these groups. Overall, there were four different tests performed, with two levels each. The data was first sorted an analyzed according to the nine thematic sets in Table 2. After this, the data was then sorted by threatening vs. non-threatening uses. Collocation analysis were run on each of the two sets, with frequency counts and MI scores recorded for both. Though the tests were performed on all sets, for the most part, frequency counts were the focus of the threatening vs. non-threatening set, and MI scores the thematic set. This emphasis was due to the purpose of each set, and the relevance of the statistical analyses for each. Though MI scores and frequency counts both provide some information about structure and some about topics being discussed, MI scores tended to be more useful for the thematic sets and frequency counts for the threatening vs. non-threatening. The thematic sets formed the basis of the discourse analysis, showing in what linguistic contexts users were employing each of the themes. The purpose of the threatening vs. non-threatening categorization, then, was to look at how this context-based variation was expressed in larger patterns in the linguistic structure of the tweets. The frequency counts were then more useful for identifying these overarching trends, whose significance was informed by the findings from the thematic set discourse analysis.

Though any MI score over 3.0 is usually considered statistically significant, because the sets are small, the cutoff was raised to 4.0 for this study. Below the 4.0 cutoff, common function words that would not generally be considered meaningful, such as “a” and “the,” tended to appear. Additionally, in order to ensure that MI scores were indeed representing repeated patterns, the minimum frequency count was set at 5. Overall, 172 statistically significant collocates were determined for the non-threatening categories (sports, names, adjectives, metaphors, failures, compound nouns, historical/political), and 120 for the threatening categories (Bomb-Related Nouns, Literal Nouns, Literal Verbs). The complete set would be too long to include here, but the words with the top-three MI scores for each category are listed below as a sample, sorted by whether the categories are predominantly threatening vs. non-threatening:

Table 3. MI scores for non-threatening categories, high to low

Word	Category	MI Score	Example
pictures	Historical/Political	6.222537	"old Ohio State Football pictures found"
belly	Compound Nouns	6.1221104	"belly bomb"
queen	Adjectives	5.817531	"bomb black woman"
hydrogen	Historical/Political	5.807500	"threw a hydrogen bomb someplace"
recipe	Names	5.559640	"Irish Car Bomb cocktail recipe"
belli	Sports	5.490003	"Belli Bomb"
guy	Historical/Political	5.485772	"will do everything to stop these bad guys"
javy	Sports	5.431109	"Javy bomb. Please, thank you"
cocktail	Names	5.407637	"Irish Car Bomb cocktail recipe"
called	Sports	5.390468	"called that Babe Ruth bomb"
chips	Adjectives	5.232551	"lays chips were bomb"
cry	Names	5.221224	"Hear My Cry by Khrisko Bomb"
nachos	Adjectives	4.969517	"some bomb vegetarian nachos"
truth	Metaphors	4.960769	"a major truth bomb will be fantastic"
drinking	Compound Nouns	4.85807	"drinking my juice bomb"
drops	Metaphors	4.845292	"he drops the F bomb on TV"
nuclear	Metaphors	4.768919	"like the destruction of a nuclear bomb"
dive	Failure	4.646117	"20 minutes after you dive bomb"
will	Failure	4.646117	"it will bomb"
bath	Compound Nouns	4.536142	"a bath bomb and shampoo for us"

exam	Failure	4.453472	"if I bomb this horrible exam"

Table 4. MI scores for threatening categories, high to low

Word	Category	MI Score	Example
stop	Literal Verbs	6.213899	"stop saying that, we'll bomb them"
gas	Literal Verbs	5.798861	"can make more mustard gas easy"
why	Literal Verbs	5.798861	"why didn't you bomb the parade"
fire	Literal Nouns	5.72746	"detonated bombs and opened fire"
strike	Literal Nouns	5.60461	"bomb in Yemen school bus strike"
guy	Literal Nouns	5.60461	"guy tried to get on a bus"
please	Bomb-Related Nouns	4.99257	"please tell them how you make a nuke"
called	Bomb-Related Nouns	4.71596	"called in a threat"
car	Bomb-Related Nouns	4.669305	"planned an ISIS car bomb attack"

The frequency counts were then taken for the consolidated threatening and non-threatening sets. 539 collocates were identified for the non-threatening group and 283 for the threatening one. As with the MI scores, the frequency counts are too large to include in full, but the first 30 tokens for both categories have been included below:

Table 5. Frequency counts for threatening and non-threatening sets

Threatening		***Non-Threatening***	
Word	**Frequency**	**Word**	**Frequency**
a	340	a	841
threat	270	the	518
to	209	to	408
in	197	and	322
the	185	i	249
at	94	in	217
and	93	with	217
squad	81	for	174

was	75	is	168
car	67	of	164
is	62	was	152
for	60	my	148
of	57	you	144
on	57	on	142
school	56	that	136
after	50	it	121
with	48	this	110
they	40	at	103
i	33	just	95
evacuated	32	some	95
called	30	so	89
an	29	be	83
it	29	from	83
not	28	are	80
you	28	yard	77
there	27	ass	75
this	27	off	74
man	26	like	72
that	26	me	71

As we would expect, the frequency counts share more tokens across the threatening and non-threatening sets than the MI scores ("a," for instance), although there are a few common entries even with the MI scores – the appearance of "guy" in both Historical/Political and Literal Nouns, for example, as well as "called" in Sports and Bomb-Related Nouns. For the frequency counts, much of the top-30 list is made up of function words, which are generally the most common compared to content words. Yet, we do find some variation in this list, both in function and content words. "Some" and "so," for instance, appear in the non-threatening list but not the threatening. Likewise, "squad" and "evacuated" only occur in the in the threatening list. Indeed, "evacuated" is unique not just to the top-30 list, but to the set as a whole, never appearing the non-threatening group once.

5. Discussion

5.1. *MI Scores and Category Context*

Generally speaking, we find that the MI scores reflect content-based differences among the categories. That MI scores tend to pick out the most rare words in small datasets can pose a challenge of accurate measures of significance. However, when combined with a minimum frequency count to minimize individual variation, the scores can also show us topics being discussed that are unique to each set. Therefore, the highest-scoring words often present us with a way of grounding the context of the definitions in the language itself. The context in which a definition is used is informed by the content of each category, and this content is in turn is related to their threat level.

For some categories, the collocates indicated by the MI scores have clear connections to their categories, as we can see in Table 6:

Table 6. High MI Score words by category (Literal Noun, Bomb-Related Noun, Historical/Political Bombs), with examples

Category	Word	Example
Literal Noun/Bomb-Related Nouns	pipe package	"former KSU student arrested, pipe bomb found in backpack" "package bombs killing folks in Austin Texas"
	suicide driverless roadside	"ISIS detonates 7 car bombs and a suicide bomb" "State car bomb attack using driverless vehicles" "after a roadside bomb exploded near his town"
	blast explosion fire	"Canadian soldiers killed in roadside bomb blast" "where a French explosion killed one person" "looks like smoke from a fire from a bomb attack"
	office fau austin	"SWAT team getting called to the office in Springfield" "bomb threat at FAU right before commencement" "car bombs in austin"
	teen soldiers authorities	"teen arrested for making bomb threat" "five Iranian soldiers were killed and 19 wounded" "authorities pursue 5 pipe bombs, destroy 2"
	allegedly evacuated investigates	"student allegedly put his friend and & 10-year old son" "Greendale mall evacuated after bomb scare" "bomb squad investigates suspicious package and the truck"
Historical/Political Bombs	pictures	"old Ohio State Football pictures found"
	atomic hydrogen nuclear	"Harry Truman after declaring atomic bombs" "most democrats think it was a hydrogen bomb" "liberals don't get nuclear weapons"

yemen	"attempts in Yemen, with US backing, has been a disaster"
iran	"did you tell them your backup plan? Bomb Iran"

With the Historical/Political Category, the highest-rated MI score is for "pictures." This finding makes sense in terms of the historical aspect of the category, with users often referencing photos found depicting old bombs (generally in reference to military service or presidential decisions as well, thereby connecting it to the political side of the definition). Additionally, we find that many of the other words in the set, including "nuclear," "atomic," and "hydrogen," describe different types of bombs that have been historically and politically important. Indeed, as we can see in the table, users often referring to events such as the bombing of Hiroshima (overseen by Harry Truman).

We can compare the types of bombs we see in the Historical/Political set to those the collocates in the Literal Nouns category refer to. Mentions of bombs in this category tend to be talking about much smaller-scale devices, such as "package" or "pipe" bombs. Furthermore, there are more terms describing how a specific attack was carried out, like "suicide," "driverless," and "roadside." Indeed, the Literal Noun category, as well as the Bomb-Related Noun category, appear to have many more collocates referring to the direct, immediate impact of bomb attacks. This is shown by the use of such words as "fire," "strike," "blast," and "explosion." They also tend to name more locations of attack, from more general places like "mall," "office," "school," "courthouse," and "country" to exact ones like "ksu" (Kentucky State University), "walmart," "fau" (Florida Atlantic University), "dnc" (Democratic National Convention), "boston," and "austin."

This pattern might also explain why we see a few more country names in the Literal Noun category than in the Historical/Political, a trend we might expect to see reversed. Many instances of countries in the Literal Noun category, including mentions of "yemen" and "iran," come from news articles reporting on international bombings. The Historical/Political set, on the other hand, tends to talk about people and countries in a more abstract sense, with the most direct comments referring to specific policies recently enacted and only occasionally to recent events. Therefore, we see far fewer actual places named. That being said, when specific locations do occur in the Historical/Political set, they do so with a much greater frequency than in the Literal Noun set. This is generally because a country is being used as a well-known and widely-discussed example to comment on a larger political issue, such as the bombings in Syria and foreign policy in the United States.

Furthermore, both the Bomb-Related Noun and Literal Noun categories are also much more likely to cite particular roles in bombing incidents. We see words like “teen,” “professor,” “man,” and “suspect” describing those who may have set off the bombs. “Soldiers” designates the victims of attacks, as do the place names already cited. We also see references made to law enforcement officials, such as “security” and “authorities,” who respond in the aftermath of an attack. All of this can likely be explained by the high percentage of news articles in these two categories, with the reporters attempting to create a coherent narrative surrounding the bombings. The fact that there are many instances of words common to legal or reporting conventions, including “investigates/investigating,” “allegedly,” “charged,” and “evacuated/evacuation” could support this trend. Overall, then, we find that differences in event proximity and specificity in the Historical/Political Bombs, the Literal Nouns, and the Bomb-Related Nouns sets are reflected in the MI score trends, and these trends are connected to the form and content of the tweets.

5.2 *MI Scores and Threat Level*

As we saw in the set-based comparison in the previous section, the further we move away figuratively from specific attacks, the more harmless the collocates become and the likelihood that the category is non-threatening increases. We find a clear demonstration of this pattern when we consider threat-based organization in addition to thematic organization, specifically looking at shared words between the non-threatening and threatening sets. Among the non-threatening categories, for instance, one of the most common contexts appears to be that of eating and drinking, as we can see in Table 8:

Table 8. High MI Score words by category (Adjective, Compound Nouns, Names), with examples

Category	Word	Example
Adjectives	chips Olive Garden meal	"lays chips were bomb" "Olive Garden for dinner sounds bomb" "then a bomb ass meal"
	selfies makeup looks	"and they take some bomb selfies" "I have some bomb makeup looks in mind" "makeup looks bomb"
	you're	"you're a bomb lady"
Compound Nouns	drinking juice bomb	"flavor bomb – drinking 10,000 Miles to Paradise" "juice bomb – drinking a Fountain of Youth"
	bath bomb photo bomb	"a bath bomb and shampoo for us to enjoy today" "another photo bomb"
Names	recipe cocktail Irish Car Bomb	"Irish Car Bomb cocktail recipe" "Irish Car Bomb cocktail recipe" "the ingredients of an Irish Car Bomb"

Out of the top five collocates for the Adjective group, three are related to food: "chips," "nachos," and "cookies." "Olive Garden," "pizza" "tacos," and "meal" also make appearances. For Compound Nouns, we have "drinking" and "juice," and the top two collocates for Names are "recipe" and "cocktail." Yet, though each uses these common contexts, they are used in a slightly different way in each. "Bomb" used as an adjective describes something positive, so instances in this set are usually describing food as tasting good or reacting favorably to the promise of food. Compound Nouns rely more on more metaphorical characteristics of bombs, with a "juice bomb" exploding in one's mouth. The Names uses are similar to the metaphorical, with drinks such as an "Irish Car Bomb" likening the dropping of a shot of Irish cream and whisky into a glass of stout to the dropping of a bomb.

As shown in Table 8, appearance, and self-care also seem to major themes among the non-threatening sets, particularly in the Adjective and Compound Noun groups. In the Adjective category, we see "selfies" and "makeup" as highly rated terms, as well as "looks" and "looking." These are usually in terms of compliments, as evidenced by the appearance of "you're" on the list. For the Compound Nouns category, the focus is on specific products ("bath bombs") or activities ("photo bomb"). In either case, we see a similar pattern to the food trend. The Compound Nouns have more explicit metaphorical ties to characteristics of a bomb – a bath bomb "detonates" in water, and a photo bomb may be described in terms of an ambush. With the Adjective category, however, "bomb" is used only in terms of its adopted meaning of describing something positive.

So far, we have seen the threat level expressed through differing collocates. As we might note from Table 3, however, occasionally high-ranking collocates are shared between threatening and non-threatening sets. An example of this is "called," which appears in two such opposite sets: Bomb-Related Nouns and Sports. The former category has some of the strongest ties to threatening uses, while the latter has the most harmless. If we were solely looking at MI scores alone, that this token is shared by such differing sets might lead us to dismiss it as irrelevant. However, a closer analysis reveals that that "called" has two different but important uses in each of the sets:

Table 7. "Called" by category (Sports, Bomb-Related Nouns), with examples

Category	Word	Example

Sports	called	"I just called that Mike Trout bomb" "called that Kershaw bomb" "I just called Scherzer's bomb"
Bomb-Related Nouns	called	"Greendale DMV evacuated after bomb squad called" "someone called in bomb threats to my office" "another bomb threat has been called into Northfield mall"

For Bomb-Related Nouns, "called" tends to be used most in the context of contacting bomb squads to deal with a potential threat or alerting someone of a bomb threat. In the sports set, however, it refers to "calling," or predicting the outcome of, a play. Therefore, even shared collocates reflect the unique context of their category, and in turn provide us with insight into how the various definitions of "bomb" are used.

5.3. *Metaphor and Semantic Shift*

The differences between the use of "bomb" in the Adjective and Compound Noun categories, as well as in the Sports and Bomb-Related Nouns categories, illustrates a common trend among the sets: that of varying directness, often characterized by the use of metaphor. Looking at the most threatening definitions, we find them to be the most concrete of all the categories. A defining feature of these categories is that they refer to actual bombs. Furthermore, not only are these uses explicitly situated in a specific time and location, but it is one that occurs either directly before, during, or directly after a bombing or bomb threat has taken place. We can contrast this with the Historical/Political uses, which, though similarly referring to an actual bomb and exact moment of time, are temporally and spatially distant from the event's occurrence.

In fact, as we transition into the non-threatening uses, the distance from actual bomb-related events increases. Indeed, in many ways, the ties to bombs themselves begins to weaken. We see this starting with the Compound Nouns. The bombs referenced by these words are more connected to a generalized idea of bombs more than anything else, referencing associated characteristics like explosions, detonations, or attacks. By the time we reach the Adjective category, however, these bomb characteristics seem to have mostly been lost. Though “bomb” may have originally retained some of the links that the compound noun has, such as the idea of something having impact, that sense is no longer clear in most uses; it is much more difficult to see how “cake sounds hella bomb rn” indexes the features that “bath bomb” does.

Therefore, one way to organize the categories could be on a continuum of more concrete uses to more metaphorical. The order would be as such: Literal Nouns, Literal Verbs, Bomb-Related Nouns, Historical/Political, Explicit Metaphors, Compound Nouns, Names, Failure, Sports, Adjectives, as seen in Table 9:

Table 9. Concrete to Metaphorical/Abstract Scale

Concrete			**Metaphorical**					**Abstract**	
Literal Nouns	Literal Verbs	Bomb-Related Nouns	Historical/Political Bombs	Explicit Metaphors	Compound Nouns	Names	Failure	Sports	Adjectives

On the concrete end of the continuum, Literal nouns to Bomb-Related nouns, we have words with clear, concrete ties to actual bombs. The metaphorical middle, from Explicit Metaphors to Names contain uses that are more metaphorical, as the name suggests, but where the comparison to bombs can be easily derived. Finally, with the abstract end, containing Sports and Adjectives we find uses that are so far from the tradition definition of “bomb” that that the original connection to bombs as explosive devices is no longer readily apparent. On the other hand, could argue that because Failure belongs in the final category, as it does not necessarily index any specific characteristics of an explosion, but that it still has links to a sense of catastrophe could be enough to warrant a place towards the middle of the spectrum.

This move towards abstract and metaphorical senses of "bomb" is in line with what we expect to see from semantic shift. Semantic shift refers to a change in a word's meaning that takes place over time (Wolfram & Schilling, 2015). Many times, the shift involves metaphorical or figurative extension, where aspects of the word's original meaning are applied to a new set of items based on shared characteristics. An example of this is the name of a "submarine" sandwich rising out of a likeness in shape to the under-water boat from which it gets its name (Wolfram & Schilling, 2015). If we assume that the metaphorical and abstract uses of "bomb" are the newer ones, then, semantic shift would explain their development; characteristics of the explosive device were applied to new items (such as a "bath bomb"), and then were gradually metaphorically extended more and more until they reached the point of complete abstraction.

5.4. *MI Scores and Structure*

Though the focus of the MI analysis was pinpointing topics associated with different uses of "bomb," we can also see how certain patterns in linguistic environment are starting to take shape. Many of the words with high MI scores that we have observed such as "pictures" so far demonstrate how key features (e.g. time) of each category are displayed through their most unique or significant words. That these words are good indicators of the content each tweet contains therefore suggests their ability to separate threatening from non-threatening tweets. While many of these collocates so far have been most helpful at determining situations in which the definitions are often used, they can also point us toward certain formulaic phrases unique to each category that could be used to distinguish them. To demonstrate this phenomenon, we can look at two groups whose heavy reliance on context seems to make them the least likely candidates for this sort of categorization: sports and names.

To some degree, the MI scores can help us measure how context-dependent certain definitions are – that is, how strongly their patterns of use rely on specific situations or people for which coding might be difficult. In the sports set, 20 out of the 39 collocates chosen as significant are players' names, for example. The names section, of course, presents a similar situation. Though there are terms that seem as though they might be more general, such as "cry," they often are part of place or song names ("Hear My Cry"). Indeed, out of the 26 significant collocates, 15 are part of names. We can see this tendency towards name-based tokens in Table 10:

Table 10. Name-based tokens with a High MI score by category (Sports, Names)

Category	Tokens
Sports	belli, javy, aguilar, tucker, yonder, yadi, harper, shaw, yo, poutler, ohtani, travis, stanton, tebow, tim, cespedes, yelich, acuna, judge, bruce
Names	cry, hear, khrisko, louder, irish, car, cherry, bangtan, factory, taylor, swift, iran, gap, stop, dropped

These context-heavy trends appear to present a problem for corpus-based categorization software. Firstly, it would be time consuming to enter in every single sports players' name into the program as a way of sorting out sports- or names-based tweets. Adding onto this is the fact that there is no guarantee that every name would be captured; in fact, while the highest-rated names in the sports set are major league players, much of the list appears to be made up of players from club sport teams, for which no centralized database exists. Therefore, if these were the only distinguishing characteristics of the sports and names categories, it would be nearly impossible to account for them in the software as is.

We find structural patterns even in these groups, however. In the sports group, the phrases "run" and "yard" are marked as statistically significant collocates of "bomb", and present ways of differentiating it from more potentially threatening usages. Both terms describe specific plays in each of the sports, with "run" coming from baseball and "yard" from football. Therefore, they provide us with specific formulas – e.g. "a [number]-run/yard bomb" – that would be unusual to find in other thematic groups. Likewise, in the names group we find the term "playing," as in "now playing [song title]." Again, "now playing" only appears in the context of song names, and as such could be one means of distinguishing this definition.

Table 11. Structure-based MI Score words by category (Sports, Names, Adjectives), with examples

Category	Word	Example
Sports	run yard	"a monster 2nd inning, 5-run bomb at Barclays" "@patrickmahomes5 threw a 30 yard bomb"
Names	playing	"now playing Hear my cry by Khrisko Bomb!"
Adjectives	rn as/asf omg	"cheese sounds really bomb rn" "them before class naps were bomb af" "just made trash juice and omg it's bomb"

We also see similar patterns in some of the other non-threatening, seemingly context-dependent definitions. As shown in Table 11, the adjective category has a large number of acronyms common to computer-mediated communication in the MI scores list, including "rn" (right now), "af/asf" (as fuck), and "omg" (oh my God). Perhaps surprisingly, phrases like "bomb rn" or "bomb af" are limited to the non-threatening tweets, and more specifically, to the adjective group. They appear to be directly tied to the casual and informal tone of the adjectival tweets. Therefore, taken together, the findings from the Adjectives, Sports, and Names categories could help make up a larger matrix of features associated with exclusively non-threatening uses. Given this trend, we see that even in the most context-dependent definitions of "bomb", we are already able to make predictions about its linguistic environment based on the MI scores.

5.5. *Frequency and Structure*

We find the patterns from the MI scores, as well as a few new ones, further reflected in the frequency counts. For the non-threatening tweets, many of the significant results come from the adjective category. One explanation of this could be the size of the set, as the adjective set is by far the largest out of all the categories. Even taking this into account, however, this set also has the most distinct tone out of all the non-threatening tweets, containing exclusively friendly and casual conversation between users. Other categories, particularly Metaphors and Compound Nouns, contain these informal examples, but also include brand advertisements or tweets with a more critical tone. Therefore, much of what we find is similar to the findings for the acronym category, discussed above, with the structure containing patterns that reflect the tweet's casual and often familiar tone. We see some instances of this in Table 12:

Table 12. High frequency count words among non-threatening categories, with examples

Category	Word	Example
Adjectives	so	"I'm educated n so bomb at what I do"
	-ass	"thx for being a bomb ass math teacher"
	some	"I just took some bomb selfies in my bathroom"
Metaphors	like	"like a bomb ticking down towards detonation"

The first pattern is that the adjective group is more likely to contain intensifiers, particularly "so" ("I'm college educated and so bomb at my job"). The phrase "-ass," as in "bomb ass selfies," also falls into this category. This trend makes sense; we would expect to see more intensifiers here both because they are more frequent in informal, everyday speech and because they are often used to strengthen adjectives. Secondly, it appears that this form of "bomb" is also more common with certain determiners, particularly "some." Though "some" does appear in other sets, as we would expect from a common function word, the phrase "some bomb" only appears in the adjective category. Furthermore, it often appears in conjunction with the other features noted, such as intensifiers, e.g. "some bomb ass Huevos Rancheros."

Therefore, because these phrases only appear in the adjective set, we could use them to potentially identify non-threatening tweets.

Thirdly, some of the most frequent words support the idea that non-threatening uses are more likely to be metaphorical. One of the highest-ranked words, for example, is “like.” Using a KWIC analysis, we see that “like” is often used to make metaphorical comparisons to bombs, as in “hit the deck like a bomb was incoming.” Furthermore, even when “like” is being used in a less comparative sense, it still remains tied to the non-threatening categories. For instance, it is often used to provide examples of recent events in the historical/political category, such as “I don’t care if its mass stabbings like China.” As such, “like,” particularly in the absence of other threatening features, could be another means to identify non-threatening tweets.

Finding patterns among the threatening tweets is a bit more challenging[2]. This is mainly because there are far fewer direct threats in the set than there are reports about threatening events. With the few direct threats that do exist, then, it can be hard to establish broader patterns, as there is not enough data to work with compared to the other uses. That being said, however, there are a few trends we can find, represented in Table 13:

Table 13. High frequency count words among threatening vs. non-threatening sets, with examples

Set	Word	Example
Non-Threatening	(a)bout to	“my hair bout to be bomb asf”
Threatening	gonna	“I’m gonna bomb my school I hate everyone here”
	evacuated arrested suspicious	“Greendale mall evacuated after bomb scare” “man arrested for building a bomb” “suspicious device prompts bomb squad response”

The first is that of “gonna” versus “about to.” “Gonna” is one of the few verbs that appears in most of the direct threats we see, with “I’m gonna bomb my school” being a notable

[2] It should be noted that for the purposes of this paper, the term “threat” includes identifications or descriptions of unsafe situations that people in the area might want to avoid.

example. While it does appear in the non-threatening tweets as well, it could signal a threat when used in combination with other threatening features. These could include the words identified in the MI scores section, as well as more generally threatening phrases like “kill” or references to weapons. Additionally, the phrase “about to” (as well as variants such as “bout to”) only appears in the non-threatening corpus, generally in the more informal senses characterized by the adjective set. Both “gonna” and “about to” represent actions intended to be carried out in the near future, but their difference in use could signal a difference in pragmatic use and suggest another way of sorting tweets.

Finally, a subset of the threatening tweets serves the purpose of alerting users to harmful situations. These are often much easier to identify, as they are more likely to follow certain genre conventions. Indeed, many of these tweets come from news sources providing story updates over Twitter. The news conventions we find include certain lexical items that are uncommon in everyday speech, such “evacuated,” “arrested,” and “suspicious.” This type of tweet may also include certain formulaic phrases or constructions such as “due to” (“due to bomb threat and courthouse evacuation”). These conventions allow us to pinpoint not only threatening tweets, but this specific subset – those tweets relating to dangerous events.

6. Conclusion

Overall, not only do we see the content of the different thematic categories reflected in their word choice in the tweet, but these observations also allow us to begin to differentiate threatening tweets from non-threatening ones. Through MI scores, we are able to find evidence of the contexts in which each of the ten identified categories are used, which include news articles and personal statements of political opinion. Based on these contexts, we also are able to observe a potential semantic shift occurring with the word “bomb” by way of metaphorical extension, giving way to new definitions, and drawing us further away from the “bomb” as a deadly weapon. Furthermore, through genre conventions and formulaic phrases, the MI scores also allow us to make initial observations about the structure of threatening tweets. The frequency counts then support these findings, as well as suggest new correlations, such as those between adjectival uses of “bomb” and intensifiers.

The sample size for this study was rather small – this was mainly out of necessity, as performing discourse analysis on larger set would not have been feasible. That being said, one possible avenue for future research could be to look at a corpus with more tokens, particularly with the help of machine learning software that had been trained on the categories discovered in this study. Furthermore, a comparison could be performed against other corpora of American English, particularly spoken ones, to see how closely the use of “bomb” in computer-mediated communication matched that of face-to-face, everyday speech. Future studies could also take into account demographic data, as available, to get a better sense of how different communities are using “bomb,” particularly looking at age, location, and race/ethnicity.

Works Cited

Bednarek, M. & Caple H. (2014). Why do news values matter? Towards a new methodological framework for analyzing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society, 25*(2), 135-158. doi: 10.1177/0957926513516041

Bock, J.G. (2012). *The Technology of Nonviolence*. Cambridge, MA: MIT Press.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139-173.

Clark, I. & Grieve, J. (2017). Dimensions of abusive language on Twitter. *Proceedings of the First Workshop on Abusive Language Online*, 1-10.

Davies, M. (2018-). *The 14 Billion Word iWeb Corpus.* Available online at https://corpus.byu.edu/iWeb/.

Eisenstein, J. (2015). Systemic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics 19*(2), 161-188. doi: 10.1111/josl.12119

Faiz, R. (2006). Identifying relevant sentences in news articles for event information extraction. *International Journal of Computer Processing of Oriental Languages, 19*(1), 1-19.

Frazier, F. & Koo, H. (2019). Discourse-pragmatic functions of the present perfect in American English TV and radio interviews. *Text & Talk, 39*(1), 77-98. doi: 10.1515/text-2018-2019

Krennmayr, T. (2015). What corpus linguistics can tell us about metaphor use in newspaper texts. *Journalism Studies, 16*(4), 530-546.

Matheson, D. (2018). The performance of publicness in social media: tracing patterns in tweets after a disaster. *Media, Culture, & Society, 40*(4), 584-599. doi: 10.1177/0163443717741356

Muresan, S., Gonzalez-Ibanez, R., Ghosh, D., & Wacholder, N. (2015). Identification of nonliteral language in social media: A case study on sarcasm. *Journal of the Association for Information Science and Technology, 67*(11), 2725-2737.

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2018). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications, 117*, 139-147. doi: 10.1016/j.eswa.2018.08.044

Salama, A.H.Y. (2011). Ideological collocation and the recontextualization of Wahhabi-Saudi Islam post-9/11: a synergy of corpus linguistics and critical discourse analysis. *Discourse and Society, 22*(3), 315-342. doi: 10.1177/0957926510395445

Semino, E. (2017). Corpus linguistics and metaphor. In Dancygier, B. (ed.) *The Cambridge Handbook of Cognitive Linguistics.* Cambridge, MA: Cambridge University Press, 463-476.

Simpson, L (2017). #OrdinaryMeaning: Using Twitter as a corpus in statutory analysis. *Brigham Young University Law Review, 2*, 487-524.

Small G. S., Medsker, L. (2013). Review of information extraction technologies and applications. *Neural Computing and Applications, 25*(3-4), 533-548. doi: 10.1007/s00521-013-1516-6

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., … Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE, 13*(10), 1-22. doi: 10.1371/journal.pone.0203794

Virtanen, T. (2009). Discourse linguistics meets corpus linguistics: theoretical and methodological issues in the troubled relationship. *Language and Computers*, *69*(1), 49-56.

Wolfram, W. & Schilling, N. (2015). *American English: Dialects and Variation*. Hoboken, NJ: Wiley-Blackwell.