Identifying Warning Behaviors of Violent Lone Offenders in Written Communication

Lisa Kaati and Amendra Shrestha
Department of Inf. Technology
Uppsala University
Uppsala
Email: firstname.lastname@it.uu.se

Tony Sardella
Washington University
St. Louis, MO
USA
Email: sardella@wustl.com

***Abstract*—Violent lone offenders such as school shooters and lone actor terrorists pose a threat to the modern society but since they act alone or with minimal help form others they are very difficult to detect. Previous research has shown that violent lone offenders show signs of certain psychological warning behaviors that can be viewed as indicators of an increasing or accelerating risk of committing targeted violence. In this work, we use a machine learning approach to identify potential violent lone offenders based on their written communication. The aim of this work is to capture psychological warning behaviors in written text and identify texts written by violent lone offenders. We use a set of features that are psychologically meaningful based on the different categories in the text analysis tool Linguistic Inquiry and Word Count (LIWC). Our study only contains a small number of known perpetrators and their written communication but the results are promising and there are many interesting directions for future work in this area.**

I. Introduction

Threats from violent lone offenders such as mass murders, isolated terrorists, and school shooters, pose a serious security threat to society and due to the many recent attacks conducted by violent lone offenders this threat has attracted increasing attention from policymakers as well as law enforcement. The most difficult part in detecting a violent lone offender is that they can come in any size, shape, and ethnicity and represent any ideology [1]. The lack of a common profile or modus operandi among violent offenders increases the challenge to identify, target and arrest persons that act on their own [2]. Another difficulty with lone offender is that since they act alone or with minimal help form others communications within a group can not be intercepted. As noted by Baaker and Graaf the challenge to prevent violent lone offenders is enormous and any counter terrorism response can only partly reduce this threat or limit the impact [2] .

In some cases violent lone offenders signal their upcoming attack prior to an event [3]. Notifications have occurred in various forms of social media or through the release of a manifesto. One example of a lone offender that signaled an attack before it took place is Dylan Roof that killed nine persons in a church shooting in Charleston, South Carolina. Before the attack, Roof published a manifesto on his own website where he stated his views towards black people. Another example is Anders Behring Breivik that before his attack in Norway 2011, wrote a Manifesto called ”2083: A European Declaration of Independence” which he distributed to approximately 8000 people via email.

Even though far from all violent lone offenders signal upcoming attacks a potential way to counter and prevent future attacks could be to detect signals in written communication that might proceed an attack. This kind of signals can be seen as weak signals that can be combined and analyzed as described in [4]. By combining several different weak signals more information about the intention, the capability and the opportunity that a violent lone offender have to commit an attack. There are several different weak signals that can provide such information. For example: signs of an individual’s radical beliefs and extreme hate, knowledge about how to produce homemade explosives, or interest in firearms and signs of rehearsal (explosive or shooting). Apart from more concrete signals, it is also possible to search for more complex signals such as different forms of warning behaviors [5].

In this work, we investigate how machine learning techniques can be used to identify psychological warning behaviors in written communication. The warning behaviors we are interested in identifying are behaviors that indicates that the writer has intention to engage in targeted violent acts. Our experiments are done on a small dataset consisting of texts written by violent lone offenders including school shooters, mass murderers and lone actor terrorists. Even though the dataset used in our experiments is a small sample, we believe that some conclusions can be drawn from the results and that computerized methods for identifying warning behaviors in written communication could be an important part of a threat assessment when trying to detect violent lone offenders before they commit an act of targeted violence.

A. Outline

This paper is structured as follows. Section II describes psychological warning behaviors and linguistic indicators as well as some work that is related to ours. Section III describes the features we have used to used to train a model to recognize texts written by violent lone offenders. Section IV describes the setup and the datasets that we have used in our experiments and Section V the results of our experiments. Section VI contains a discussion of the results. Finally, we conclude the

TABLE I
THE DIFFERENT CATEGORIES PRESENT IN LIWC

Category	Abbreviation	Examples	No of words	Category	Abbreviation	Examples	No of words
Word count	WC			Female references	female	girl, her, mom	124
Summary Language				Male references	male	boy, his, dad	116
Variables				Cognitive processes	cogproc	cause, know, ought	797
Analytical thinking	Analytic			Insight	insight	think, know	259
Clout	Clout			Causation	cause	because, effect	135
Authentic	Authentic			Discrepancy	discrep	should, would	83
Emotional tone	Tone			Tentative	tentat	maybe, perhaps	178
Words per sentence	WPS			Certainty	certain	always, never	113
Words $>$ 6 letters	Sixltr			Differentiation	differ	hasn't, but, else	81
Dictionary words	Dic			Perceptual processes	percept	look, heard, feeling	436
Linguistic Dimensions				See	see	view, saw, seen	126
Total function words	funct	it, to, no, very	491	Hear	hear	listen, hearing	93
Total pronouns	pronoun	I, them, itself	153	Feel	feel	feels, touch	128
Personal pronouns	ppron	I, them, her	93	Biological processes	bio	eat, blood, pain	748
1st pers singular	i	I, me, mine	24	Body	body	cheek, hands, spit	215
1st pers plural	we	we, us, our	12	Health	health	clinic, flu, pill	294
2nd person	you	you, your, thou	30	Sexual	sexual	horny, love, incest	131
3rd pers singular	shehe	she, her, him	17	Ingestion	ingest	dish, eat, pizza	184
3rd pers plural	they	they, their, they'd	11	**Time orientations**	TimeOrient		
Impersonal pronouns	ipron	it, it's, those	59	Past focus	focuspast	ago, did, talked	341
Articles	article	a, an, the	3	Present focus	focuspresent	today, is, now	424
Prepositions	prep	to, with, above	74	Relativity	relativ	area, bend, exit	974
Auxiliary verbs	auxverb	am, will, have	141	Motion	motion	arrive, car, go	325
Common Adverbs	adverb	very, really	140	Space	space	down, in, thin	360
Conjunctions	conj	and, but, whereas	43	Time	time	end, until, season	310
Negations	negate	no, not, never	62	**Personal concerns**			
Other Grammar				Work	work	job, majors, xerox	444
Common verbs	verb	eat, come, carry	1000	Leisure	leisure	cook, chat, movie	296
Common adjectives	adj	free, happy, long	764	Home	home	kitchen, landlord	100
Comparisons	compare	greater, best, after	317	Religion	relig	altar, church	174
Interrogatives	interrog	how, when, what	48	Death	death	bury, coffin, kill	74
Numbers	number	second, thousand	36	**Informal language**	informal		380
Quantifiers	quant	few, many, much	77	Swear words	swear	fuck, damn, shit	131
Psychological Processes				Netspeak	netspeak	btw, lol, thx	209
Affective processes	affect	happy, cried	1393	Assent	assent	agree, OK, yes	36
Positive emotion	posemo	love, nice, sweet	620	Nonfluencies	nonflu	er, hm, umm	19
Negative emotion	negemo	hurt, ugly, nasty	744	Fillers	filler	Imean, youknow	14
Anxiety	anx	worried, fearful	116	**Drives**	drives		1103
Anger	anger	hate, kill, annoyed	230	Affiliation	affiliation	ally, friend, social	248
Sadness	sad	crying, grief, sad	136	Achievement	achieve	win, success, better	213
Social processes	social	mate, talk, they	756	Power	power	superior, bully	518
Family	family	daughter, dad, aunt	118	Reward	reward	take, prize, benefit	120
Friends	friend	buddy, neighbor	95	Risk	risk	danger, doubt	103

paper with Section VII that draws some initial conclusions and proposes some directions for future work.

II. WARNING BEHAVIORS AND LINGUISTIC INDICATORS

Although no common profile predicts a lone offender, indicators of a forthcoming attack may be observable when conducting threat assessment. A threat assessment is a set of investigative and operational activities designed to identify persons to be at risk for violence [6]. As stated in [6] one of the fundamental principles of threat assessment is that: "targeted violence is the result of an understandable and often discernible process of thinking and behavior". This indicates that acts of targeted violence not impulsive or spontaneous - something that supports that it might be possible to identify certain warning behaviors that are present in an attacker before an attack takes place. There are many different threat assessment protocols that has been tested and evaluated in different situations. In [7] a set of psychological warning behaviors that can be used as a part of a threat assessment is described. A warning behavior is any behavior that "precedes an act of targeted violence, is related to it, and may, in certain cases, predicts it".

The possibility of detecting potential perpetrators leveraging written communication from social media (blogs, forums, messaging systems) is an approach recognized in [8]. In [8] the mechanisms for ensuring that indicators of leakage are highlighted can be used as a potentially powerful tool in detecting lone-actor terrorists (including school shooters), particularly those exhibiting right-wing ideologies. However, revealing a plan or a direct threat is not what we are trying to identify in this work. Instead we are considering the possibility to detect psychological warning behaviors, something that is also described in [9]. Our work builds on the idea that warning behaviors are present in texts written by violent lone offenders but not in texts written by persons that do not engage in targeted violence. Our work investigates the possibilities for machine learning to capture warning behaviors in written texts

by using psychological meaningful features.

Previous research [10] [11] has shown that it is possible to connect word use to psychological constructs such as personality, drives, cognition and emotion. The text analysis tool Linguistic Inquiry and Word Count (LIWC) [12] sorts words in psychologically meaningful categories. By counting the relative frequencies of words in a text and dividing words into different meaningful categories it is possible to gain a deeper understanding about the person who wrote it. LIWC calculates (in percent) how much a person uses words from the different LIWC categories. One benefit with LIWC is that it can serve as a way to gain indirect information about subjects who will not directly provide information about themselves. The study described in [13] noticed that texts written by ten different lone offenders prior to their engagement in targeted violence had some common linguistic indicators that could be used as a part of a threat assessment.

As noted in [14], little research has been completed on computerized text analysis of threatening communications. However, several studies have demonstrated that word use is linked to psychological states, and the underlying psychology of a speaker or author can be revealed through text analysis [14]. Some of the warning behaviors identified by [7] have been identified as potentially observable in social media [9] and a first attempt to detect warning behaviors has previously been conducted using term dictionaries and keywords [15]. Based on the relative frequency of words from psychologically meaningful categories a set of eight possible indicators of the drives and emotions that preceded engagement in targeted violence is identified. Another approach to assess the risk of threatening communication is described in [16] where the risk of violence in threatening communications is assessed using two different methods: LIWC and topic models. Topic models are based on the statistical properties of the documents themselves rather than a set of predetermined words. Results indicated that words reflecting anger, death, or negative emotions (from LIWC) were not used more frequently in the threats documents considered. The topic modeling method indicated that some of the themes in the written communications could be related to psychotic symptoms that are characteristic of a number of disorders.

III. Psychologically Meaningful Features

Domain knowledge is imperative in order to understand what features to use in the machine learning algorithms to build an accurate model. Each feature should contain a piece of information that can be used in the creation of the predictive model. In this specific case, we are interested in capturing weak signals of a warning behavior or a psychological state that a person shows before committing an act of targeted violence. In order to capture the psychological state, we have used the different categories from the text analysis tool Linguistic Inquiry and Word Count (LIWC) [12]. Using LIWC categories as features, we hope to obtain data independent features and the aim is to investigate the possibilities to capture psychological warning behaviors that separates lone offenders from other social media users based on written communication.

LIWC is a computerized word counting tool that searches for approximately 4,000 words and word stems and then categorizes them into grammatical (e.g., articles, numbers, pronouns), psychological (e.g., cognitive, emotions, social), or content (e.g., achievement, death, home) categories [11], [17]. LIWC counts words in psychologically meaningful categories by counting the relative frequency of words from the different categories.

IV. Experimental Setup

In the experiments we have used Adaboost, an algorithm introduced by [18]. Classification trees were used as the base classifiers, and the R ada package was used for the implementation. The R ada package provided means for estimating the optimal number of base classifiers. The R ada estimation was used as the number of base classifiers.

A. Dataset

The dataset that we use in our experiments includes subjects who have alone, or with minimal help from others, perpetrated a violent act with the intent to cause severe (fatal) personal damage. Prior to or in relation to their act, the individuals have stated their points of view in written communication that has subsequently been made public, either by themselves or by police investigations. The whole dataset consists of a combination of school shooters, ideologically motivated offenders and mass murderers. The dataset consists of a total of 32 subjects. Some of the subjects have written several texts, resulting in a dataset with a total of 46 different texts. All texts in the dataset are assumed to have been written by the subjects. The texts from the school shooters were collected from [19], a database that has a compendium of documents relating to a wide range of active shooter incidents in educational settings. We have used the texts that are described as the "Shooters' words".

The written communication from the ideologically motivated offenders and mass murderers were collected from publicly available sources and consist of written communication or manifestos. No single ideology can be defined among the included subjects, rather their ideologies and motivations are widely spread. The subjects that are ideologically motivated offenders and mass murderers that are included in this study and some information about their texts can be found in Table II.

In order to represent texts that are not considered to be written by violent lone offenders , we have used three other datasets. First, we used a set of 54 blogs that were publicly available [20]. Most of the blogs were written about personal interests, news, fashion and photography. This dataset is denoted as $Blogs$. We have also used a set of users and their posts from the web forum named Stormfront. Stormfront was founded in 1995 and is commonly described as the leading white supremacist web forum. The forum has grown into what may be the Western world's most popular forum for

TABLE II
STATISTICS ON THE SUBJECTS' TEXTS.

Name	Number of words	Publication year	Age when writing
Nidal Malik Hasan	3555	2008	38
James von Brunn	47178	1999	79
Anders Behring Breivik	807712	2011	32
Dylan Roof	2446	2015	21
Elliot Rodger	108206	2014	22
Christopher Dorner	11489	2013	34
Jim David Adkisson	1056	2008	58
Ted Kaczynski	34719	1995	52
Lucas Helder	3296	2002	21
Andrew Joseph Stack III	3236	2010	54

TABLE III
THE DIFFERENT DATASETS USED IN THE EXPERIMENTS.

Dataset	**No of instances**	**Description**
LO	46	Texts written by violent lone offenders
$Blogs$	54	Blogs about personal interests, news, fashion and photography.
$Stormfront$	108	Posts written by a set of Stormfront users.
$Boards$	108	Posts written by users from the Irish forum Boards.ie.

racists, Holocaust deniers and criminals to post articles, engage in discussions and share news of upcoming racist events [21]. Data from Stormfront users is denoted as $Stromfront$. Stormfront is a hate inspired web forum where discussions were often highly negative while the blogs more commonly focused on personal interests and were more positive in nature. In a report published in 2015 [21], Stormfront members are held responsible for murdering almost 100 people. An example of a registered active member is the Norwegian Anders Bering Breivik who murdered 69 people in a rampage that aimed at stopping what he saw as a Muslim invasion of his country.

Finally, we have included a data from the Irish discussion board Boards.ie (https://www.boards.ie/). Boards is public forum where things like hobbies, politics and sports is discussed. Data from Boards.ie users is denoted as $Boards$. The $Boards$ dataset consist of the posts from 108 users that posted more than 100 posts and less than 200. We only use text that can be assumed to be written by the user, all citations are removed. The reason for choosing three such different datasets to represent texts that are not written by violent lone offenders is that we want to capture representative picture of social media where violent lone offenders might publish texts prior to an attack. However, more social media data from different forums and networks should be included to make the experiments more realistic.

Table III describes the different datasets we have used in our experiments. In order to get an understanding of the different datasets, we have extracted the 10 most frequently used words in each dataset after removing function words (words that have little lexical meaning). The results are shown in Table IV. The Blogs and Boards.ie users use words like *get, go, think, people* and *more*. Not surprisingly, many words relating to ideology were among the most frequent words from the texts from Stormfront, for example *jews, race, hitler* and *white*. The texts written by the violent lone offender frequently use some of the words that were used in the Blogs and Boards.ie: *all, more, any* and some unique words *europe, muslim* and *western*. However, since the texts are very different in size we can not draw any conclusions about it the frequent use of these words is representative for each individual text. We can only conclude that these words are frequently used words in the different datasets.

V. RESULTS

Four experiments were conducted on the different datasets in order to gain an understanding on the possibilities of using machine learning to identify warning behaviors that preceded a violent attack present in text. Due to the small sample size, we used Leave-One-Out Cross-Validation (LOOCV) [22] on the model as a way to overcome the small sample size. The model was built leaving out a single sample, which was later used to derive a prediction for the left-out sample. The small training dataset was balanced since classification models were not prepared to cope with unbalanced data set since they under perform when the data is unbalanced [23]. Synthetic samples of the minority class were generated using Synthetic Minority Over-sampling Technique (SMOTE) [24]. The SMOTE technique is used on the training set only within the loop and tests are done on remaining single record. This insures that we are not introducing bias in the evaluation of the model. The algorithm selects two or more similar instances using a distance measure and resamples an instance's attribute individually by a random amount within the difference to it's neighbor.

We have used all LIWC categories as features in our experiments; however, not all features necessarily contributed to the classification. Therefore, we used Mahalanobis distance [25] as a mechanism to rank features and select the optimal group that improves the classification results. The most important features used in the classification in the different experiments are shown in Table V.

The results from the experiments we have conducted is reported using confusion matrices in which we present the

TABLE IV
THE 10 MOST FREQUENTLY USED WORDS IN THE DIFFERENT DATASETS WITHOUT FUNCTION WORDS

Dataset	**Top 10 words**
LO - texts written by lone activists	all, more, one, people, europe, any, many, muslims, western, life
$Blogs$ - personal blogs	all, one, good, get, see, got, make, more, much, go
$Stormfront$ - Stormfront forum users	people, white, more, one, any, think, national, race, jews, hitler
$Boards$ - Boards.ie users	all, get, one, think, any, good, people, more, go, got

TABLE V
THE MOST IMPORTANT FEATURES USED IN THE CLASSIFICATION IN THE DIFFERENT EXPERIMENTS

Experiment	**Most important features (LIWC categories)**
Experiment 1	Article, Personal pronouns, Negative emotions, See, Total pronouns, Prepositions, Anger, Differentiation, Affective and Perceptual processes
Experiment 2	Perceptual processes, Prepositions, Negative emotion, Anger, Function words, Time, Informal language, Relative, Certain, Impersonal pronouns, Leisure and 3rd person plural
Experiment 3	Time, Articles, Personal pronouns, See, Differentiation, Prepositions, Quantifiers, Negative emotion, Biological processes, Cognitive processes
Experiment 4	Assent, Articles, See, Negative emotion, Differentiation, Auxiliary verbs, Personal pronouns, Anger, Social, Affective processes

number of true positives, false negatives, true negatives, and false positives as illustrated in Table VI.

The measures we use to report our results are sensitivity which is defined as:

$$\frac{TP}{TP+FN}$$

precision defined as:

$$\frac{TP}{TP+FP}$$

and recall defined as:

$$\frac{TP}{TP+FN}$$

We also use the measures accuracy and balanced accuracy. Balanced accuracy is used to avoid inflated estimates on imbalanced datasets. Accuracy is defined as:

$$\frac{TP+TN}{TP+FP+TN+FN}$$

and balanced accuracy as:

$$\frac{0.5 \cdot TP}{TP+FN} + \frac{0.5 \cdot TN}{TN+FP}$$

We have done four different experiments to get an understanding on the possibilities to detect texts written by violent lone offenders. One of the largest challenges when working with the kind of machine learning techniques that we are is to find a representative opposite class. To gain a deeper understanding of we vary the opposite class in the experiments.

TABLE VI
CONFUSION MATRIX

	Predicted class	
Actual class	**True Neg. (TN)**	**False Pos. (FP)**
	False Neg. (FN)	**True Pos. (TP)**

A. Experiment 1

In the first experiment the goal was to separate the texts written by lone offenders from other texts. The opposite class was created by combing the blogs, the texts from the Stormfront forum and the texts from Boards.ie. When using all LIWC categories as features, 13 of the 46 violent lone offender texts were incorrectly classified and 16 of the 270 texts from the blogs and forums were incorrectly classified. In an attempt to improve the results, we used Mahalanobis distance to rank the features. The most important features were: Article, Personal pronouns, Negative emotions, See, Total pronouns, Prepositions, Anger, Differentiation, Affective and Perceptual processes. The same experiment was done using only the 11 most important features which improved the results to 35 correctly classified texts written by lone offenders (out of 46) and 242 correctly classified texts from the blogs and forums (out of 270). These results are reported in Table VII .

B. Experiment 2

In experiment 2, the goal was to separate texts written by violent lone offenders from the blogs that were written about personal interests, news, fashion and photography. The results are shown in Table VIII. As can be noted, 7 (out of 46) lone offender texts were incorrectly classified and 4 (out of 54) bloggers were incorrectly classified.

The most important features for the classification were *Perceptual processes, Prepositions, Negative emotion, Anger, Function words, Time, Informal language, Relative, Certain, Impersonal pronouns, Leisure* and *3rd person plural*. The some example words from each category can be found in Table I.

C. Experiment 3

In experiment 3, the goal was to separate texts written by violent lone offenders from texts written by Stormfront users. The results are shown in Table IX. Out of the 46 lone offender

TABLE VII
RESULTS FOR EXPERIMENT 1

	$Blogs+$ $Stromfront+$ $Boards$	LO	**Accuracy**	**Specificity**	**Recall**	**Precision**	**Balanced Accuracy**
LO	242	11	0.8766	0.7609	0.8963	0.9565	0.8286
	28	35					

TABLE VIII
RESULTS FOR EXPERIMENT 2

	$Blog$	LO	**Accuracy**	**Specificity**	**Recall**	**Precision**	**Balanced Accuracy**
LO	50	7	0.89	0.8478	0.9259	0.8772	0.8869
	4	39					

TABLE IX
RESULTS FOR EXPERIMENT 3

	$Forum$	LO	**Accuracy**	**Specificity**	**Recall**	**Precision**	**Balanced Accuracy**
LO	100	7	0.9026	0.8478	0.9259	0.9346	0.8869
	8	39					

TABLE X
RESULTS FOR EXPERIMENT 4

	$Boards$	LO	**Accuracy**	**Specificity**	**Recall**	**Precision**	**Balanced Accuracy**
LO	100	4	0.9221	0.9130	0.9259	0.9615	0.9195
	8	42					

texts, 7 were misclassified and out of the 108 texts from Stormfront 8, were incorrectly classified.

The most important features in the classification were *Time, Articles, Personal pronouns, See, Differentiation, Prepositions, Quantifiers, Negative emotion, Biological processes* and *Cognitive processes*.

D. Experiment 4

In experiment 4, the goal was to separate texts written by violent lone offenders from texts written by Boards.ie users. The results are shown in Table X. Out of 108 Boards.ie user texts 8 were incorrectly classified and out of the 46 lone offender texts 4 were incorrectly classified. The most important features in the classification were *Assent, Articles, See, Negative emotion, Differentiation, Auxiliary verbs, Personal pronouns, Anger, Social* and *Affective processes*

VI. DISCUSSION

The goal of this work was to investigate if machine learning can be used to identify texts written by violent lone offenders. Since there seems to a possibility to detect certain psychological warning behaviors in written communication it might be the case using psychologically meaningful features and machine learning, it is possible to identify texts written by violent lone offenders before they engage in targeted violence.

We have only used the LIWC categories as features in our experiments since we wanted to use only data independent features and not features that depended heavily on our dataset. This is particularly important in cases where there are small unbalanced datasets as in our case. The results from the different experiments and the ranking of the features provides some interesting perspectives on common linguistic markers among the lone offenders.

The first experiment attempted to separate texts written by violent lone offenders from blogs, texts written by Stormfront forum users and texts written by Boards.ie users. The results showed that using only the 11 most important features we obtained an accuracy of 0.8766, a specificity of 0.7609 and a sensitivity of 0.8963. The features that were used for the

classification shows that the use of pronouns and personal pronouns played an important role in the classification. Two other linguistic dimensions that played an important role in the classification: Articles (such as a, an, the) and Prepositions (e.g $to, with,$ and $above$). Another feature that played an important role in the classification is negative emotions (in particular anger that is a subcategory to negative emotions in LIWC). The rest of the features that played an important role in the classification are various psychological processess. These processes (corresponding to categories in LIWC) are Perceptual processes (e.g. $look, heard, feeling$), See (e.g. $view, saw, seen$), Differentiation (e.g. $hasn't, but,$ and $else$), Affective processes (e.g. $happy, cried$) and Biological processes (e.g. $eat, blood, pain$).

In the second experiment we tried to separate texts written by violent lone offenders from bloggers. The linguistic dimensions that played an important role in the classification were Prepositions, function words, impersonal pronouns, personal pronouns and third person plural. Time orientations (Time and Relativity) and Personal concerns (Leisure, Death, informal language) also played an important role. Psychological processes that played an important role were Perceptual processes (e.g. $look, heard,$ and $feeling$), Tentative (e.g. $maybe, perhaps$), Certainty (e.g. $always, never$), along with negative emotions and anger. Further investigation on how these psychological processes can be related to previous theories on warning behavior still needs to be conducted.

In experiment 3, we attempted to separate texts written by violent lone offenders from texts written by Stormfront forum users. The features that played an important role for the classification are shown in Table V. The linguistic dimensions that played an important role in the classification were Article, Quantifiers, Prepositions and Personal pronouns. The psychological process that played an important role in the classification were Differentiation, See, Biological processes, Cognitive processes and Negative emotion.

In experiment 4 we tried to separate texts written by violent lone offenders from texts written by Boards.ie users. The linguistic dimensions that played an important role in the classification were Article, Auxiliary verbs and Personal pronouns. Personal concerns in the form of Assent and psychological processes such as Differentiation, See, Affective processes and social processes. Negative emotions and anger also played an important role in the classification.

Our experiments are done on three different data sources (and one experiment where all the data is combined). One of the goals with using the different data sources is to get a deeper understanding on what separates the texts included in the different sources from the texts written by violent lone offenders. The linguistic dimensions that plays an important role in all experiments are personal pronouns and prepositions. The use of pronouns in natural language has been studied previously and the use of pronouns have been linked to different aspects of personality and emotion [26]. For example can a frequent use of third person plural ($they, them$ etc) in a group suggest that the group is defining itself to a large degree by the existence of an oppositional group [17]. Why prepositions (e.g words like $to, with,$ and $above$) are important features needs to be investigated further.

Negative emotions also played an important role in the classification, in particular anger (experiment 1, 2 and 4). When Pennebaker and Chung [17] studied al Qaida-texts they discovered that texts were relatively high in emotion and that the relation between positive and negative emotion differed from what is common in natural conversation. Natural conversation usually contains almost twice as many positive than negative emotion words while the al Qaida-texts had a much higher relative degree of negative emotion words, mostly anger words [17]. One important feature when separating texts from Blogs and Stormfront from the texts written by the violent lone offenders is anger (with anger words such as $hate, kill$ and $annoyed$).

The psychological process Differentiation were important in the classification except in experiment 2 where the Blogs are used as an opposite class. Differentiation is a category in LIWC where people are distinguishing between entities, people, or ideas, including words such as $but, without,$ and $hasn't$. Perceptual processes (e.g $observing, heard$ and $feeling$) were also important in the classification. In particular the category See (experiment 1, 3, and 4) that contains words such as $view, saw,$ and $seen$.

Features about personal concerns only mattered when separating blogs and Boards.ie texts. This might reflect the fact that personal concerns are expressed more on the Blogs and Boards.ie than on Strormfront forums where ideology is discussed more.

VII. Conclusion and Future Work

Even though our dataset is small we believe that our experiments shows a possibility to use machine learning to identify texts written by violent lone offenders. Even if we do not know if warning behaviors plays a role in the classification, we can conclude that some psychological categories from LIWC are different when comparing "regular" texts and texts written by violent lone offenders. We believe that with additional research, computerized support for analyzing written communication can be used as a part of a manual threat assessment. However, more experiments need to be conducted with these kind of techniques prior to applying them as a part of larger system to detect potential violent lone offenders. The warning behavior that we are trying to identify in this work relates to targeted violence, but it might be possible to use a similar approach to detect signs of other warning behaviors, such as suicidal behavior. A direction for future work would be to do other forms of analysis - both computerized statistical and qualitative on the texts in our dataset. A qualitative analysis could give some indication on whether any previously defined psychological warning behaviors (such as the eight different warning behaviors described in [7]) are present in the texts while a computerized text analysis could give us some additional information on similarities and dissimilarities between the different texts.

Automatically monitoring the Internet for potential lone offenders in the name of security raises a new set of questions. There are many ethical issues and issues relating to the personal integrity to be considered before using automated techniques to detect warning behaviors. We view automated techniques as an aid for human analysts in their evaluation and search for potential threats towards the security of the society.

REFERENCES

[1] L. Kaati and P. Svenson, "Analysis of competing hypothesis for investigating lone wolf terrorist." in *European Intelligence and Security Informatics Conference (EISIC)2011*, 2011.

[2] E. Baaker and B. de Graf, "Preventing lone wolf terrorism: Some ct approaches addressed," *Perspectives on Terroris*, December 2011.

[3] P. Gill, J. Horgan, and P. Deckert, "Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists," *Journal of Forensic Sciences*, vol. 59, no. 2, p. 425435, 2014.

[4] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson, "Harvesting and analysis of weak signals for detecting lone wolf terrorists," *Security Informatics*, vol. 2, no. 1, pp. 1–15, 2013.

[5] M. Fredholm, *Understanding Lone Actor Terrorism. Past Experience, Future Outlook, and Response Strategies*. Routledge, 2016.

[6] R. Borum, R. Fein, B. Vossekuil, and J. Berglund, "Threat assessment: Defining an approach for evaluating risk of targeted violence," *Behavioral Sciences and the Law*, vol. 17, pp. 323–337, 1999.

[7] J. Reid Meloy, J. Hoffmann, A. Guldimann, and D. James, "The role of warning behaviors in threat assessment: An exploration and suggested typology," *Behavioral Sciences & the Law*, vol. 30, no. 3, pp. 256–279, 2012.

[8] C. Ellis, R. Pantucci, J. de Roy van Zuijdewijn, E. Bakker, B. Gomis, S. Palombi, and M. Smith, "Lone-actor terrorism final report," *Countering Lone-Actor Terrorism Series No. 11*, 2016.

[9] K. Cohen, F. Johansson, L. Kaati, and J. Clausen Mork, "Detecting linguistic markers for radical violence in social media," *Terrorism and Political Violence*, 2013.

[10] J. Pennebaker, M. Mehl, and K. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[11] Y. Tausczik and P. J.W., "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, March 2010.

[12] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count (liwc): A text analysis program." New York: Erlbaum Publishers, 2001.

[13] K. Cohen, L. Kaati, and A. Shresta, "Linguistic analysis of lone offender manifestos," *International Conference on CyberCrime and Computer Forensics (ICCCF)*, 2016.

[14] J. W. Pennebaker and C. K. Chung, "Using computerized text analysis to assess threatening communications and actual behavior." in *Threatening communications and behavior: Perspectives in the pursuit of public figures*. The National Academies Press, 2011.

[15] F. Johansson, L. Kaati, and M. Sahlgren, "Detecting linguistic markers of violent extremism in online environments," *Combating Violent Extremism and Radicalization in the Digital Era*, pp. 374–390, 2016.

[16] K. Glasgow and R. Schouten, "Assessing violence risk in threatening communications," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 38–45.

[17] J. W. Pennebaker and C. K. Chung, "Computerized text analysis of al-Qaeda transcripts," in *The Content Analysis Reader*, K. Krippendorf and M. A. Bock, Eds. Sage, 2008.

[18] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *In Proceedings of the thirteenth International Conference on Machine Learning.*, 1996, pp. 148–156.

[19] "School shooters .info," Last accessed: 2016. [Online]. Available: https://schoolshooters.info/

[20] "Google blogs." [Online]. Available: https://www.blogger.com/

[21] H. Beirich, "Intelligence report," 2015. [Online]. Available: https://www.splcenter.org/fighting-hate/intelligence-report/2015/20-years-hate

[22] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[23] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[25] P. Mahalanobis, *On tests and measures of group divergence I. Theoretical formulae*. J. and Proc. Asiat. Soc. of Bengal, 1930.

[26] J. W. Pennebaker, *The secret life of pronouns: What our words say about us.* CT New York: Bloomsbury Press., 2011.