`

The wisdom of the crowd: Twitter as a tool to predict violent events.

Prepared for: Natali Dolgova, Ph.D.

GWU

EP6111

Research Report

Prepared by: Lilian Sao De Rivera

December 17, 2017

Table of Contents

List of Figures

1. Introduction

It has been a long journey since the appearance of the concept "Wisdom of the Crowds" for the first time in 1907, and the use of Collective Predictive Analysis based on information gathered from the web. The concept was first used by Sir Francis Galton, an English statistician who lived around 1822 and 1911. Sir Galton made calculations on the values that were provided by farmers of the area about the correct weight of an ox in a local fair. The observation showed that the average of the guesses was approximately the same value as the correct weight of the animal (Galton, 1907). Sir Galton concluded that the correct guess was based on the expertise of the crowd: all of them were farmers that worked with cattle and knew by personal experience the weight of these animals. This research is presented as a baseline by Surowiecky (2004) to address the importance of crowds and shared opinion to solve problems. Surowiecki (2004) discuss how the wisdom of the crowds has shaped business, economies, societies, and nations. But, he states that not all crowds generate wisdom. In order for a crowd to be wise, it needs to comply with certain characteristics: diversity, independence, decentralization, aggregation (p. 10).

Figure 1. The graphic is based on Surowiecki (2004) discussion about the characteristics of a group.

Diversity is necessary to maintain a varied range of views, knowledge, and approaches to solve problems. Independence in a group means that each member makes his or her own decision without being influenced by any member of the group. This characteristic is very important to balance the wisdom in the group, and the lack of this quality may lead to decisions in the group that cannot be labeled as wise. This fact is due to the likelihood of increasing the probabilities of

making the same mistakes because the members of the groups have the same beliefs: they have been influenced. Decentralization in a group means that each member is capable of acquiring knowledge locally. Lastly, aggregation is the capability to gather the opinions of everyone in the group to reach a solution, a conclusion or a vote (Surowieki, 2004). The wisdom of the crowd can be used to solve three types of problems: cognition problems, coordination problems, and cooperation problems. Cognition problems need a definite solution, for example, to determine who is going to win the world cup in 2018. Coordination problems look into the possibility of collaboration between the members of the group to accomplish a main goal, knowing that everybody is trying to do the same activity; for example, Waze is an application used to avoid traffic and is fed by members of an adjacent community. Cooperation problems refer to the type of cooperation that members of the group seek to solve a problem; for example, crowdfunding that is used to finance ventures.

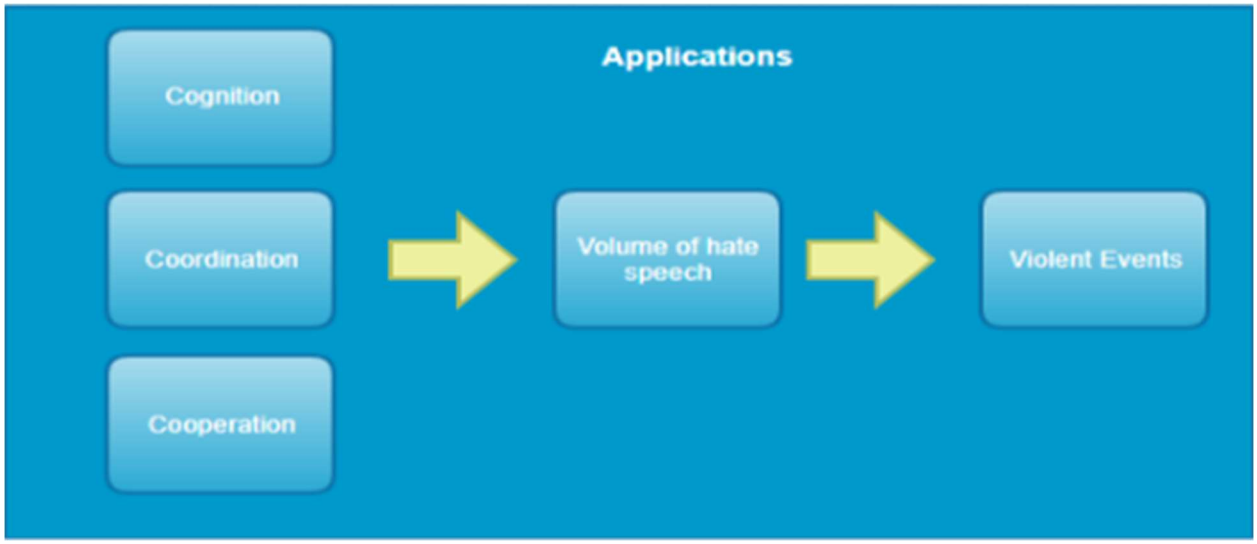

Figure 2. Types of problems approached by wisdom of the crowds and where violent events fall into these categories.

The present study addresses malicious behavior. Malicious behavior is a problem of coordination, if we follow this categorization. But this is not a problem that generates smart decisions. On the contrary, the communication that is generated from the members of the groups -hate speech- generates malicious behavior, like riots or hate crime. These actions fall into the category of coordination, people in the group are organized to perform a pernicious action. That is because the group lacks an important factor: independence. The group is highly motivated and influenced by members of the group.

Then, we may assume that malicious behavior could be a result of the different interrelations between groups that share a final goal. This goal is to express opinions with actions that are influenced by hate speech that is spread in the neighborhood by a communication medium. The communication medium is social networks which is an example of the aggregation characteristic

described by Surowieki(2014). There have been studies about the predictability of violent events after a highly publicized event (Burnap & Williams, 2015). These studies have stated that after the event triggers an outburst of social media communication, this communication is spread through the neighborhood of the event. This vicious communication can generate violent behavior like a hate crime against a specific community, riots, and other types of violent activities. An example of this type of social behavior can be seen in the 9/11 terrorist attacks. After this event, an increase in hate crime against Muslim communities was observed. Most recently, we could mention the death or Freddy Gray in custody of Baltimore's police, which sparked violent riots in the city of Baltimore. Burnap and Williams (2015) researched the possibility of predicting the violent events. The elements that were studied were the amount of hate speech in social media, the event, and the location of the event. These researchers generated a learning machine based on hate speech around the murder of Lee Rigby in London by Islamic extremists. The model proved to be successful, according to the study. Burnap and Williams worked in building a corpus associated with hate speech. The corpus was tested with a service on Internet that provides human assurance of the context of the work. Burnap and Williams (2015) established a set or rules in order to create a learning machine that could learn new words and predict malicious behavior. The detailed description of the learning machine is not included in the study since it is being used by Homeland Security.

There are some studies that address the possibility of predicting malicious behavior. It is an ongoing subject that has become more feasible due to the widespread use of social networks. According to Mantzaris (2014), this new medium has generated abundant investigation that can involve several aspects of human behavior. Social networks have opened the door to people to express themselves in ways that were not possible before. This situation generates opportunities to analyze such behaviors. If these behaviors could be translated into predictive models that can be monitored by people with decision power, then hate crimes, riots, among other types of malicious actions against communities, could be avoided or minimized by acting based on protocols of security designed for each type of event. Even though there are data scientists working on this subject, there is room to contribute to the field due to the novelty of the topic and the new techniques of analysis in big data that were developed in recent years. The creation of technology to manage high quantities of information like in-memory computing and the creation of databases like Hadoop, Elastisearch, MongoDB amongst others. The introduction of deep learning and learning machines that have allowed the monitoring of sentiment in social networks and identify images and speech could only be accomplished by a human being. The ability of machines to interpret written language and read context is the interest of this study. The research is looking to discover a relation between hate speech and pernicious actions in the vicinity of the neighborhood without a widely publicized event. The main objective is to discover patterns of ongoing activity associated with discontent or unstable environments. This analysis is important to make public policy to regulate and prevent such events since it could produce clear paths of actions.

In order to accomplish the task at hand, two data sets are being considered: hate speech of six months from the Twitter feed in South Sudan, and data from GDELT database (Global Database of Events, Language, and Tone). GDELT was created by Dr. Kalev Leetarau. It comprises

information around the world based on the news, social media, video among other types of publications. Twitter is used based on the recommendations from the papers by Manzartis (2014) and Burnap and Williams (2015). According to these authors, Twitter is the media that most people use to express opinions and share the news. Once the information is gathered from these two sources of information, a scatter plot will be generated in order to determine the type regression model to be used to generate the predictive model. If we find a significant correlation between hate speech and malicious behavior we can generate a predictive model to forecast possible pernicious actions. Then, the model could be tested based on the appropriate correlation associated with the chosen model, but first, it is important to find if there is an important correlation and a relationship of cause an effect between the two variables.

The structure of the research is as follows: first, a methods section will explain the different steps that were executed during the research; second, the results of the methods are presented; and finally, the conclusions and results are given.

2. Methods

The research needed to delimit the size of the information in order to comply with the schedule proposed to deliver the results. The following parameters were chosen to attain the desirable outcome:

- An African country: the study needed an unstable environment.
- Twitter was used to detect malicious speech: as recommended by the studies.
- The information from the country was used without dividing the territory: this vicinity was useful to detect the possibility of predictability.
- The corpus used to detect hate speech is composed by the following words: coward, jieng, MTN, nuer nuer, and nyam myam.

It was established that two datasets were required the volume of instability in the region and the volume of hate speech. The datasets were collected via API from GDELT database and the API of the Twitter feed. This information was cleansed and prepared to be processed into graphics. The graphics were used as a previous analysis to observe visually if there could be a possibility of predictability. Two sets of graphics were elaborated: the first set was to observe the variables involved in the study, that is to observe the movement of events over time and compare both lines. If the lines in the time indicated some correlation visually then the second graphics would indicate if there could be a correlation. The type of graphic more appropriate for this task is scattered plots. Because time is a variable important in the study, there were included four graphics to observe the visual relationship between the two variables over a time period of four weeks.

2.1 Datasets

One of the datasets was acquired via the API GDELT tool (Application Programming Interface). We obtained an Excel document with the observations of 7 months, from January 2016 until July 2016. The observations depicted the number of individuals involved in a riot or a violent event. The dataset presented several observations for one day. In order to make the information more readable, the dataset was cleansed from the extra fields that were not used in the research and the observation were summarized to present one record per day and per week. The Fig.1 below shows the behavior of the variable in a period of 31 weeks.

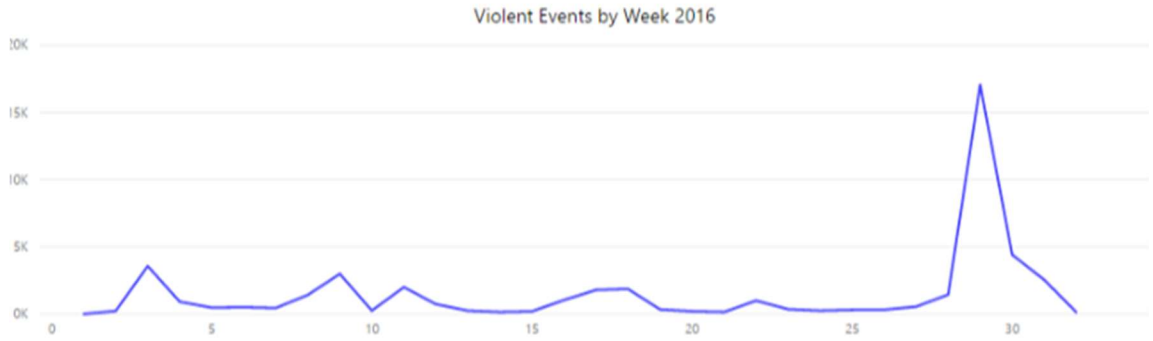

Figure 3. This graphic represents the violent events summarized by week from the GDELT database.

The other dataset was obtained via the Tweeter feed. Because words associated with a specific connotation were needed, that is hate speech, the feed was verified by Crimson Hexagon to comply with the initial requirements of the study. Crimson Hexagon is a commercial product that is used to analyze sentiment in words. Crimson Hexagon provided an Excel file with the five specific hate words that were used in this context: coward, jieng, MTN, nuer nuer, nyam nyam. The information provided was an observation per day per word. The time frame used was the same used by the GDELT database, so both sets of observations could relate to each other. The information was cleansed from all irrelevant data and summarized by a total of words, by day, and by week. The figure below shows the behavior of the variable in a 31 weeks period. The information depicted shows the movement detailed by word in the first graphic and the summarization of all five words is shown in the second graphic.

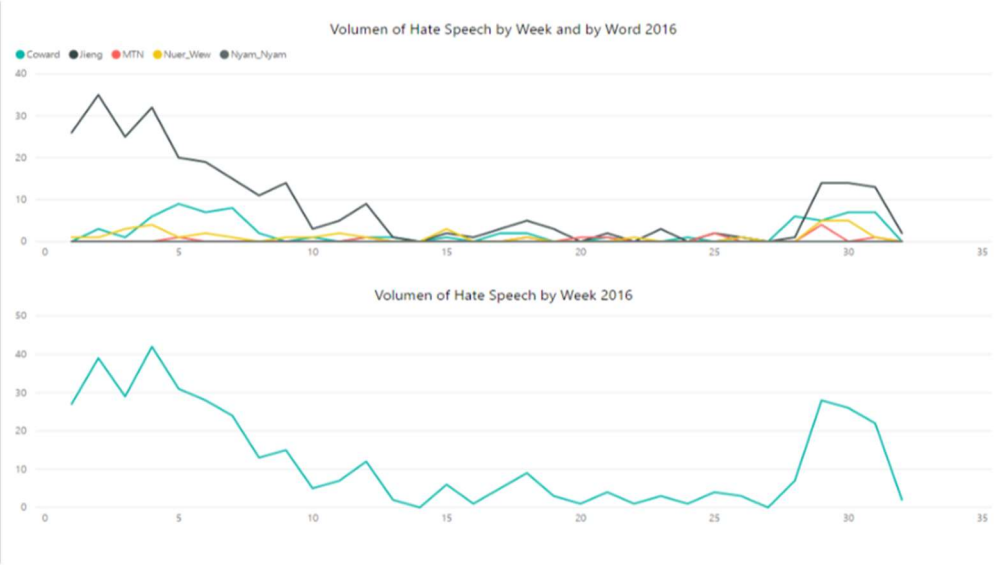

Figure 4. Graphic that represents hate speech detailed by word and summarized from the Twitter feed.

2.2 Definition of Variables

In order to determine if malicious behavior could be predicted from the volume of hate speech in an unstable environment, first, we need to determine if there is a strong correlation between the two variables. Then, if such correlation exists, a more detailed study can be made in order to generate a predictive model. In pursuit of this purpose, we proceed to define the independent variable and the dependent variable. The independent variable is a volume of hate speech since the hypothesis states that an increase in hate speech generates violent events. Hence, the dependent variable is the violent events. Because it is important to understand the type of correlation we proceed to elaborate a graphic with both variables to make a simple visual observation. The resulting graphic is shown below. From this graphic, it can be observed that there is some kind of relation between the hate words and the volume of hate speech. It can be observed also, that there are two peaks that occurred between the two variables at the same time. This observation could mean that a relationship between the two variables might be possible.

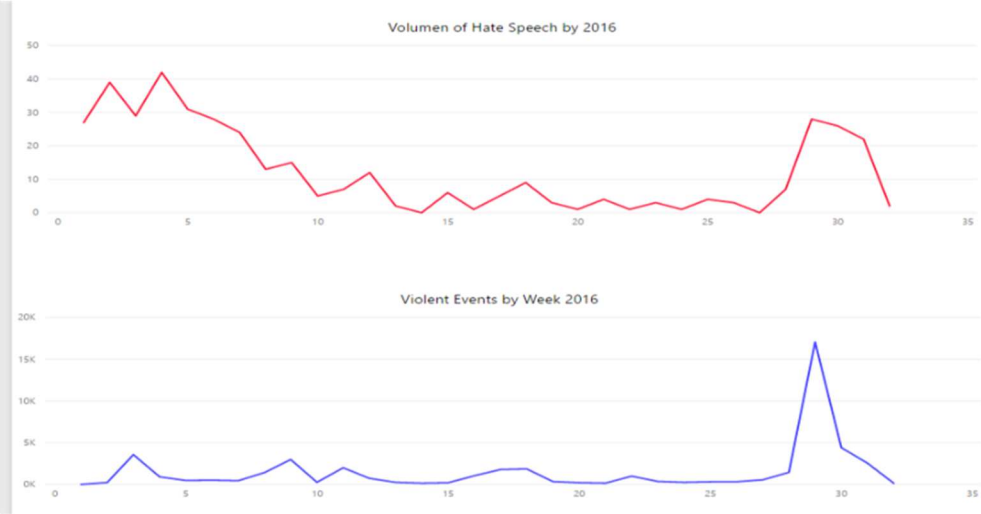

Figure 5. Independent variable hate speech from Twitter feed versus dependent variable violent events from GDELT database.

But more visual analysis should be done. Now that is clear which variable is the independent variable and the dependent variable; we need to use a scattered plot graphic But, we are adding a third variable in the study: time. Since this study is an exploratory research about the possibility of predictability; we need to address the time factor. The study searched for a short period of time. The volume of hate speech will generate violent events in the vicinity of the time that the volume of hate speech increases in social media. But we do not have a specific time frame, so it was decided to plot four weeks. The first plot below shows the hate speech versus violent events in the same week. The second graphic shows the violent events that occurred a week after hate speech, the third graphic shows an elapsed time of two weeks and the last one shows three weeks of events from the hate speech activity. The four graphics show an accumulation of dots at the beginning of the time period. Then the dots continue to disperse along a horizontal line. This dispersion seems to be more a more scattered. At the same time, the four graphics show the movement of an outlier. In the first week, the outlier is in the middle of the scatter plot and continues moving backward in each plot. This outlier was checked on the original files of information and represented only one event in the 31 weeks of analysis so it was decided to ignore this behavior. From the plots, we can deduce that there is the possibility of a horizontal line between the points, since all the dots accumulate along the x-axes of the scattered plots in all four graphics. This observation could indicate a correlation, which allows continuing with the following step: choosing the correlation method to be used with the variables.

Figure 6. Scattered plot for hate speech versus violent events with a week difference between each one.

2.3 Correlation Method

Because it can be observed that a horizontal line can be drawn in the scatter plots, the Person's coefficient method was chosen. The conditions used to choose this method are: the variables are numerical, we can see a chance in value of every pair, and the variables may have a normal distribution. The Pearson coefficient method shows the correlation between two variables and is calculated as follows:

Pxy = Cov (X,Y) / Sx * Sy

Cov(X,Y) = Covariance between hate speech and violent events.

Sx = Standard Deviation of Hate Speech

Sy = Standard Deviation of Violent Events.

Four calculations were made to determine which time is the best fit for this study: present, one, two, and three weeks. The Pearson coefficient seems the most appropriate approach because the datasets and the scattered plots seem to satisfy the conditions to use this coefficient. The results calculated in the following section will affirm or deny our initial assumption, that is, there is a correlation between hate speech and volume of violent events.

3. Results

The goal of the study as was stated in the previous section is to discover an important correlation between the independent variable -hate speech- versus the dependent variable -malicious behavior-. In order to attain the main purpose, the following scale was used to assess the results obtained from the calculations.

Table - 1
Criteria for Pearson coefficient

Strenght	Posstive	Negative
Small	.1 to 3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

According to the scale above, the results from the calculations presented in the table below can be evaluated. The ideal scenario would be to attain a large strength, but it seems at first glance that this is not possible.

Table – 2
Pearson coefficient for 4 weeks period of time

Week	Sx	Sy	Cov(X,Y)	Pxy
0	12.6924	3040.74	12006.37	0.31109
1	12.7543	3078.58	1389.569	0.03539
2	12.8566	3121.3	-4534.52	-0.113
3	12.8256	3153.73	-6402.1	-0.1583

According to the Pearson coefficient calculations above, the relation between the volume of hate speech and the victims of malicious behavior is low-medium since it has a value of 0.311. The coefficient would need at least a factor over the 0.50 to be conclusive. We can observe that with each week of elapsed time the coefficient becomes lower and lower. This fact, at least, shows a trend: the correlation between hate speech and violent speech diminishes as the weeks elapsed between these two events increases. It is interesting to notice in the scattered plot graphics a tendency for a straight line. But, the Pearson coefficient tells us that such relationship is weak. It seems that there is not enough information to calculate the correlation. The research concentrated on five hate words. It may be possible that more words would be needed to establish a good

connection. The words used could be expanded to a broader context, where casual words can be used as hate speech in spite of the meaning of the word. We could be measuring context instead of words and include more words to the corpus used in the study. During the processing of the information, one problem was detected: the dataset from hate speech did not provide location; the GDELT database does provide this information. With more accurate locations, the study could match the location of hate speech with the location of the violent events. This additional information could have provided more accurate information. The other pitfall could have been the country over which the study was made. It is necessary to obtain more information about the use of social networks in the country and the preferred microblogs that are used. Maybe the particular circumstances of the country could have misled the information. The studies of Burnap and Williams (2015), and Mantzaris (2014) rely on a high use of social networks and were performed in developed countries. This last fact can be an important factor in the parameters of the study that could have affected the research.

4. Conclusions

The main purpose of the research was to determine if there is a correlation between volume of hate speech and volume of malicious behavior. The assurance of correlation was necessary to continue with the following step: to create a predictive model. The absence of correlation makes ineffective the creation of such a model. According to the results presented in the previous section, we can make the following conclusions.

- The datasets show a .31 coefficient factor in the first week. The results show a medium-low correlation. The coefficients in the following calculations are near zero and negative which means no correlation between the variables for these weeks.
- The meaning of the low-medium correlation in the first week could mean that the volume of hate speech is caused by the violent events. This fact is the opposite of what the study intended to attain.
- The fact that the correlation diminishes its value as the time passes between the occurrence of hate speech and violent events, tells us that the time elapsed between these two events is important and that the study might yield better results based on a daily basis instead of a weekly basis.
- Twitter shows partial information according to the scatter plots. Thus, more microblogs could be added to get better results.

5. Recommendations

Because the study is not conclusive, a more thorough research could be made that addresses the problems described in the previous sections. Some adjustments could be made to obtain information that is more relevant to the type of outcome that is expected. One of the problems

found was location; both datasets need to have a location to match vicinity of events. Another problem detected was that Twitter was used as the only microblog used in the study, which could mean that important information could be missed in the study. According to Mantzaris (2014), Facebook is used to create communities. These communities seem to play an important part of hate speech because of the influence of the members of the groups that is explained amply by Surowiecki (2004). The corpus used comprised only five words of hate speech; this quantity of information may not be representative of the volume of hate speech. Lastly, an African country was used as the main source of information, a developing country. This fact could bias the information because the use of social networks might not be as widely adopted as the developing countries that were used in the studies by Mantzaris (2014) and Burnap and Williams (2015). Based on these facts, the following changes in the parameters of the research are recommended in order to retry the study:

- Narrowing the vicinity of the events and/or changing the country of the research.
- The Twitter feed needs to be completed with information from another microblog feed like Facebook.
- The study included five words that were used in the context of "hate speech". More words could be added.

6. References

Burnap, P. & Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet, 7, 223–242.* doi:10.1002/poi3.85

Galton, F. (1907). The ballot box [Letter to the editor]. *Nature*. Retrieved from http://galton.org/cgi-bin/searchImages/galton/search/essays/pages/galton-1907-ballot-box_1.htm.

Galton, F. (1907). Vox Populi. *Nature*. Retrieved from http://galton.org/essays/1900-1911/galton-1907-vox-populi.pdf

Leetarau, K. (2017). *GDELT Project*. Retrieved from https://www.gdeltproject.org/

Mantzaris, A.V.(2014). Uncovering nodes that spread information between communities in social networks. *EPJ Data Sci., 3,* 26. https://doi-org.proxygw.wrlc.org/10.1140/epjds/s13688-014-0026-9

Surowiecki, J. (2004). *The wisdom of Crowds*. New York, NJ: Random House Inc.

Twitter API. (2017). *Developer documentation*. Retrieved from https://developer.twitter.com/en/docs