IDENTIFYING SPAM IN SOCIAL LISTENING DATA

Group project for the SCC460 Data Science Fundamental course

Group 7: Oleksa Stepaniuk (32341885), Adi Karri, Mansoor Muneer Reehana (35051820), Chia-Yen Chiang (32404170)
Company: Relative Insight

Introduction

Social Networking is a fundamental part of modern day life. They provide millions of users with access to other users around the world which would be considered a pipe-dream a few decades back. These social network operate in different ways. For example Instagram and Snapchat are image and video based sharing social networking outlet whereas Twitter is mirco-blogging platform which permits user to post up to 280 characters. These service are utilized not only for personal use by individuals but for commercial use also, by companies ranging from small to large across many sectors. Companies take advantage of these social network platforms to gain better insight into their target market. These services provide a great number of benefits however they all suffer from the same issue, Spam.

There are companies such as Brandwatch which scrap social media based data however they don't remove the spam from the data-set hence the data that they provide is very much raw. This isn't the most appropriate state for the data to be in. This is where a company like Relative Insight steps in. Relative Insight is a Lancaster-based company that was established in 2012. Currently, Relative Insights helps their clients to understand what language, words, idioms people use to talk and describe different products and situations. Relative Insight clients include R/GA, Disney, Unilever, Havas, and Pearson. One of the main aims of the company is to determine whether a social media entry is spam or not.

The aim of Relative Insight is not to merely collect data but to gain insights from the vast amount of available data in today’s digital world, however much of it is plagued by spam. Spam is more prevalent in certain social media services compared to others. It's important to highlight that Relative Insight deal with text based data. Image based platforms such as Instagram, Snapchat and Flickr aren't ideal since it is difficult to extract a significant amount of text based data from these platforms. Although 'Facebook' is a text heavy platform, it is harder to obtain large amounts of data from it due to the recent privacy concerns which were raised by US Congress (Watson, 2019) For these reasons the company tends to focus more on Reddit and Twitter. Reddit does have spam however it is generally easy to identify it as it has user based voting and user tend to down vote "spam". In contrast it is much more difficult to detect spam on Twitter.

Spam is quite a subjective word that has many definitions. In a cooperation with the Relative Insight we defined spam as consisting of:

1. Irrelevant tweets
2. Spam or abusive tweets as defined by Twitter
3. Tweets of companies/professionals (e.g., bloggers, etc.)

Our research question was to build a classifier using machine learning techniques to detect spam, as given in our definition, on Twitter.

In our methodology we will discuss the literature we reviewed as part of our research strategy, as well as an in depth review of how we collected, integrated and pre-processed the data. We then apply several machine learning classifiers upon the data and further optimize the best one. Further on, in our results, we discuss the results and what our analysis revealed as well as the potential biases we encountered.

1. Methodology

The first issue when dealing with spam on SNS is what is the definition of spam. What is spam to a particular user may not be spam to another user. **Heymann etc al. (2007)** defined spam as *“either content designed to mislead or content that the sites legitimate users don’t wish to receive”*. They suggested a three pronged approach to deal with spam on Social networking sites. In Identification based (detection) approach, manual identification is performed by users or moderators of what is a spam which can serve as a training set of labels and act as prerequisite for further machine learning models. In rank based approach they suggest returning lower ranking to a possible spam, based on certain criteria, on a search result list for example a most popular search result on a SNS. In the interface based approach they suggest using CAPTCHAS’s to prevent automated account creation and to restrict access to some features on the sites for newly created accounts. In our project with relative insights labelling of what is a spam is not provided by users or moderators.

Zinman and Donath etc al. (2007) worked on spam detection on SNS like Facebook and MySpace. They directed their research particularly on detecting spam in unsolicited communications like friend requests. They differentiated between genuine unsolicited communications like a friend request from a stranger and those profiles which mimic ordinary users but are actually spam which promote commercial interest like advertising products or invitations to pornographic sites. They categorised the profiles into two dimension based on Sociability and Promotions. Now based on attributes like number of images in comments section, number of links in comments, number of unique comments of friends, networking section etc., each of the profile was scored on sociability and promotions. The results were normalised between 0 and 1. Machine learning algorithms like K nearest neighbours (KNN), back propagation neural networks and Gaussian Naïve Bayes were used to classify. However, the results were in the range of 30-50% accuracy which was not significantly better than random selection. This study provided a way forward for further research in this area. This approach relates well with our task. However much of the information like number of images in comments section, number of unique comments of friends etc. was not available in the dataset provided by the Relative Insight. We used suspended account information from Twitter which would have taken these aspects into consideration for labelling those profiles as spam.

Social bots on Twitter are becoming the new face of spam with growing record of malicious applications like manufacturing fake political support/opinions, Fake news, Promotion of terrorist propaganda and recruitment, manipulating the financial markets etc. **Varol etc al (2017),** designed a framework for detecting social bots. To train the system they used a meta data of tweets from 15K manually verified Twitter bots identified via a honey pot approach along with 16K verified human accounts. From this meta data features were extracted. User based features includes number of friends and followers, number of tweets by the user and profile description. Network features include retweets, mentions and hashtag occurrences. Temporal features like average rate of tweets over period of time were extracted along with language and sentiment features. Machine learning algorithms were used from the scikit-learn library and the model evaluated based on the Area under the receiver operating characteristics (ROC) curve with 5- fold cross validation. Random forests, Adaboost, Logistic Regression and Decision Tree classifiers were used. Random forest gave the highest performance with 95% Area under the curve. This machine learning algorithm called Botometer has been provided free access online. Overall the study estimated that 9%-15% of accounts on Twitter are social bots. We used the scikit-learn libraries and the AUC to model and evaluate our machine learning algorithms. Also honey pot approach can be used by the end client in future to provide labels for spam data so that our classification accuracy can be improved.

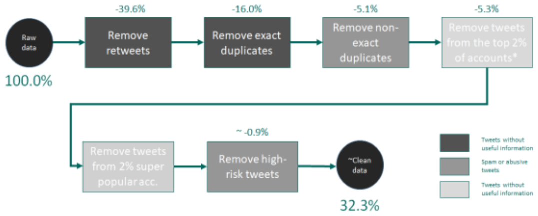
There has been continuing research in this area with many new research papers being published, however the overall underlying aspect among most of them is the extraction of various features from the social media sites, be it profile information or content based information which can provide distinctive attributes on which the Machine learning algorithms can distinguish between a spam and genuine messages.

2. Pre-processing

In the light of our definition of spam, some messages can, for example retweets, can be unambiguously defined as irrelevant tweets. However, significant share of messages can be defined as spam only with some probability. Consequently, section 2.1 describes the approach to cleaning the datasets from the unwanted tweets that can be identified with certainty and section 2.2 describes how to prepare the dataset for the development of predictive models described in later parts of the report.

2.1 Pre-processing as an answer to research question

Steps of pre-processing Twitter data are described in Picture 1:

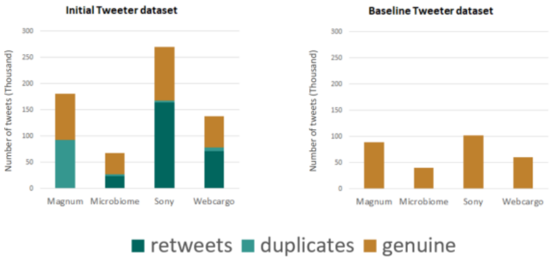

-39.6%
-16.0%
-5.1%
-5.3%
Raw data
Remove retweets
Remove exact duplicates
Remove non-exact duplicates
Remove tweets from the top 2% of accounts*
100.0%
~ -0.9%
Remove tweets from 2% super popular acc.
Remove high-risk tweets
~Clean data
32.3%
Tweets without useful information
Spam or abusive tweets
Tweets without useful information

Picture 1. Steps of cleaning the dataset
* Based on the number of tweets in the database per account after removing retweets and exact duplicates. In our case threshold is > 5 tweets

Data provided by the Relative Insight included 18 Excel files (0.74 gigabytes of information). After merging them we formed an initial dataset with 656,000 observations. Each observation includes a tweet and information about the Twitter account, for instance, number of followers. All tweets are related to one of the four subjects:

1. Ben & Jerry’s ice cream (Magnum dataset)
2. Sony playstation (Sony dataset)
3. Gut related subjects, e.g., gut bacteria, gut-healthy diet, digestive disorders, etc. (Microbiome dataset)
4. Sex drive (Webcargo dataset)

Almost 40% of the initial dataset are retweets. They are not useful for the Relative Insight purposes because they do not include original messages. Luckily retweets can be identified with absolute certainty, each retweet starts with four standard symbols “RT @”. Apart from retweets, initial dataset included 16% of exact duplicates. For some reason, dataset included the same tweets from the same accounts as several different observations. Almost half of the messages related to the Ben & Jerry’s topic were duplicates. We classified retweets and exact duplicates as belonging to the first group of unwanted tweets - tweets without useful information. Removing them is the first step in our pre-processing methodology.



Picture 2. Dataset after removing retweets and exact duplicates

According to the Relative Insight observations, one of the most widespread types of spam is so-called “non-exact” duplicates. These are messages that have identical text but include different URL or a hashtag. To detect them we processed text of all messages in the following way:

5. Change tweets text to lower-case
6. Remove mentions of other accounts
7. Remove hashtags
8. Remove links
9. Remove punctuation
10. Remove digits

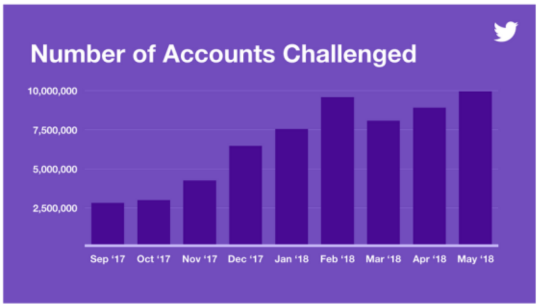
5% of tweets (as a % of the initial dataset) were identified as non-exact duplicates. Analysis of these messages confirmed that they can be regarded as spam messages and belong to the second group of unwanted messages - spam or abusive tweets as defined by Twitter.

Regarding the third category of unwanted tweets - tweets of companies/professionals (e.g. bloggers, etc.) it is not possible to identify them with certainty. Still, there are two features that are good indicators of tweets that belong to this category. Firstly, regular users usually tweet about miscellaneous topics while “professional” accounts have a narrow topic. As a result, 83% of accounts in the dataset have only one message in the dataset, while a minority of 0.5% accounts have 12 or more tweets. The absolute “champion” is account ‘NBAIndia’ that has 850 tweets mentioning Sony. While it is not possible to unambiguously distinguish professional account from a person that is passionate about a certain topic, we applied an ad hoc criterion of treating the top 2% of accounts with the largest number of tweets in the dataset as belonging to the third group. These accounts were responsible for 5.3% of tweets in the initial dataset. In a similar fashion, we identified tweets from the top 2% of the most popular accounts (with more than 23 thousand followers) as belonging to the third group of unwanted tweets.

Such an ad hoc approach is acceptable as a starting point, however, it is also possible to identify “professional” accounts using a classifier model (see Section 3.2).

2.2 Pre-processing to support model building

In the previous section, we described non-exact duplicates as belonging to the second group of spam or abusive tweets as defined by Twitter. Unfortunately, every Twitter dataset includes a large share of spam and abusive tweets that cannot be identified as non-exact duplicates. They need to be detected and removed. Even a brief analysis of a dataset consisting of tweets can convince anyone that there is no obvious way of identifying spam messages. The process of identifying spam or abusive message is inherently subjective - something that is offending or irrelevant for one group of users can be interesting and useful for the other group. Therefore the best way for identifying spam messages that are used by all social networks are user reports – if a large enough number of users report a message or an account as abusive/irrelevant, social network administrators have grounds for suspending it. Each month Twitter challenges millions of potential spam accounts (see Picture 3).



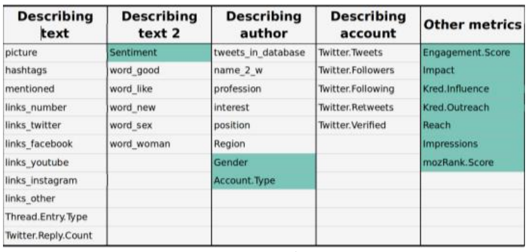
Picture 3. Number of accounts challenged by Twitter
Source: https://blog.Twitter.com/

We do not have access to the users' reports, but luckily we do not need them. If Twitter thinks that an account is a source of spam or abusive messages, it suspends it. When the page of such account is opened, the message “This account is suspended is displayed.” Thus it is possible to build a web scraper that will check which accounts in our dataset are suspended and then use it as a dependent variable to build the classifier model that will predict whether an account is or will be suspended by Twitter.

Our web scraper needs 1.5 seconds to check whether one account is suspended. Using five computers we were able to classify all 202 thousand accounts in our dataset in 15 hours. 3.6% of them (out of baseline dataset) were suspended by Twitter. Interestingly there is a strong correlation between suspended accounts and accounts that produce non-exact duplicates. While accounts that were suspended are responsible for only 4.7% of tweets in the dataset, they were the source of 21.2% of unique non-exact duplicates. This confirms our hypothesis that both groups of tweets belong to the second group of unwanted tweets - spam or abusive tweets as defined by Twitter.

The final dataset used for creating classification models includes 15,179 spam messages (produces by suspended accounts or non-exact duplicates messages) and 222,962 ‘normal’ tweets. Therefore for every spam message, we have 15 ‘normal’ messages. This is a severe case of class imbalance. Simulation studies (Weiss, 2013) showed that with the increase of class imbalance, classification error increases exponentially. To improve the situation, we divided all spam messages into the train (80% of observations) and test (20% of observations) sets. Then we randomly selected the analogous number of ‘normal’ messages and added them to each set. This is an example of using undersampling approach to avoid the class imbalance problem.

Both train and test datasets include 37 independent variables (see Table 2):



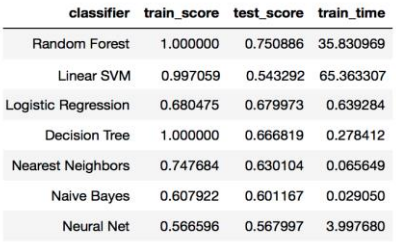
Describing text	Describing text 2	Describing author	Describing account	Other metrics
picture	Sentiment	tweets_in_database	Twitter.Tweets	Engagement.Score
hashtags	word_good	name_2_w	Twitter.Followers	Impact
mentioned	word_like	profession	Twitter.Following	Kred.Influence
links_number	word_new	interest	Twitter.Retweets	Kred.Outreach
links_twitter	word_sex	position	Twitter.Verified	Reach
links_facebook	word_woman	Region		Impressions
links_youtube		Gender		mozRank.Score
links_instagram		Account.Type		
links_other				
Thread.Entry.Type				
Twitter.Reply.Count				

Table 2. Dependent variables used for modelling

3. Results

3.1 Probability of message being spam or abusive tweet as defined by Twitter

This is an initial way of checking all the classifiers using scikit-learn. This is a quick and easy method of determining the metrics, taking no longer than a few minutes to create and execute. However this comes at a cost due to the simplicity of execution it is prone to many types of error and biases and thus, we use it only as a baseline to determine what model may be appropriate.



classifier	train_score	test_score	train_time
Random Forest	1.000000	0.750886	35.830969
Linear SVM	0.997059	0.543292	65.363307
Logistic Regression	0.680475	0.679973	0.639284
Decision Tree	1.000000	0.666819	0.278412
Nearest Neighbors	0.747684	0.630104	0.065649
Naive Bayes	0.607922	0.601167	0.029050
Neural Net	0.566596	0.567997	3.997680

Table 3. Accuracy of Multiple Classifiers on the Preprocessed dataset

The training and test data was ratio was set to 70:30.The following table shows the result of the comparison. We can see from the table that K Nearest Neighbour (KNN) and Gaussian Naive Bayes take the least training time and Linear Support Vector Machines (SVM) take the longest time to train. In terms of training scores Random Forest and Decision Tree show 100% accuracies which is a clear case of overfitting and needs to be investigated further individually by K fold cross validation. Logistic Regression seems like the best classifier overall as it provides a good balance between train time and accuracy. Furthermore we can take this and optimize it for better accuracy

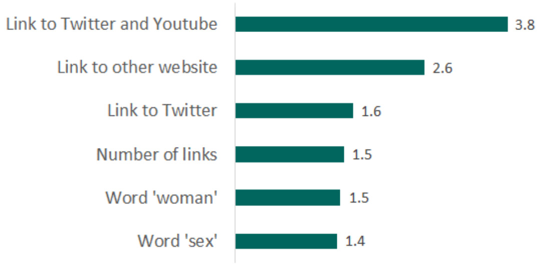
From pre-processing we have produced two dataset: one with variables provided only by the Brandwatch (variables in grey in Table 2) and one without them. As we need a model which is optimized for Brandwatch data and one for general data.

Due to numerous unique countries it was necessary to group the countries for better computation efficiency and interpretability. We grouped countries into 11 regions based on their location.

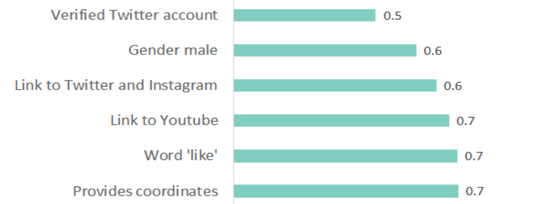
To determine a more optimal logistic model we implement a multi-directional stepwise function, based on the Akaike information criterion. From our exploratory data analysis we observed a negative relationship between the link covariates – i.e. when a user tweeted, and had a Twitter link in the tweet, the tweet was less likely to contain links to other social media platforms. This lead to the addition of interaction terms in the model to see if it would help produce a better fit.

The logistic regression based on the Brand Watch variables gave a train accuracy of the 0.702 and test accuracy of 0.703. Hence the accuracy only decrease marginally for the test data. The train recall is 0.641 and the test recall is 0.624. Hence the true positive rate has decreased for the test data.

Now for the general logistics regression, the train data accuracy came to 0.695 and the test data accuracy is 0.695. The test data accuracy has increase which is promising. The recall of the general logistic regression for the train data is 0.624 and for the test data it is 0.614. Clearly the Brandwatch model is better as it has a higher accuracy and recall value.



Picture 4. Covariates with the strongest positive effect (increase probability of message being a spam), odds ratio



Picture 5. Covariates with the strongest negative effect (decrease probability of message being a spam), odds ratio

3.2 Probability of message belonging to a company/professional blogger

In order to build a model that predicts probability of tweet being produced by an organization, public body or individual we used the same dataset as in the section 3.1, only this time the variable Account.type was used as a dependent variable. This is one of the variables provided by the Brandwatch, it indicates whether account belongs to the company or an individual.

In contrast to the previous section, we decided to try an alternative machine learning approach to building models and create a neural network. Our network has 3 layers, 36 nodes in each layer and was trained for 100 epochs. ReLU was used as an activation function.

After using resampling for improving the class imbalance problem, our model provided 89% score for the test set accuracy, precision and recall.

To check the sensitivity and robustness of the results, we used the sum of standard deviation (STD) to measure uncertainty within a model. For instance, after 5 simulations, we calculated six STD from 5 precision, 5 recall and 5 F1 scores. Then, we summed up six STD. To measure how results change when altering certain parameter, we calculated sum of percentage change for all average testing scores between default and new model. In both cases model results were not significantly affected by the changes in data and parameters.

The results of model developed in section 3.2 can be used to improve the performance of the logistic function from the section 3.1. The dataset contains a significant share of observations with missing value for the Account.type, therefore we can use labels estimated with neural network as an input data.

4. Potential Biases and Validity

Throughout the course of any project biases will arise. Particularly in research the aim should to try to reduce the bias, determine where the bias originated from. This should allow for a better evaluation of results and findings.

One of the main biases that arises due to pre-processing, is how we dealt with missing data. Due to the large amount of missing data within some features, rows containing missing data were deleted.

Labelling spam based on suspended accounts, as we did during web-scrapping, is another source of bias as the user could be suspended wrongly at that given time but later unsuspended. In several other instances, individuals have been temporarily banned/suspended due to a flaw in new updates in Twitters spam filter. Another recent example occurred when a popular meme arose which make fun of people with a liberal worldview. This involves user changing their profile to ‘NPC’. This resulted in 1500 accounts being banned even-though they weren’t spam. Here bias is two-fold, firstly as we previously mentioned, accounts may be suspended at the time we checked but later unsuspended if it was done wrongfully. In our opinion, the bias described above is unavoidable if we use suspension of the account as a spam indicator. With such methodology our model by definition cannot be better than spam filters developed by Twitter.

The bias described above is closely related to the question of the construct validity. We cannot be sure that all messages identified as spam in our dataset actually are spam and vice versa. As a result, we concentrated on the recall metrics when evaluating our model – ability to correctly identify message labeled as spam. Instances when message identified as non-spam were identified as spam are less relevant because they can be the result of problem with spam labels.

The question of internal validity (whether we can conclude that there is a causal relationship between dependent variable and covariates) of our study is complicated by the fact that we do not assume a causal relationship between variables. For example, multiple links in a tweet do not cause it to be a spam, however they can be an indicator of tweet being a spam. Therefore in our case the question of internal validity is whether our covariates represent a genuine characteristics of a spam message. We have all reasons to believe it to be true because tweet is completely defined by its text and characteristics of the account. As a result, if there exists a 'spam characteristics' they should be present in our covariates.

Lastly we tried to improve the external validity of our results by using tweets from different datasets (from Sony PlayStation to gut problems).

5. Conclusion

The aim of this project was to create a methodology for cleaning a dataset of tweets from three types of spam: irrelevant tweets, spam or abusive tweets as defined by Twitter, Tweets of companies or professionals (e.g., bloggers, etc.).

Our analysis of dataset showed that irrelevant tweets consist of retweets and exact duplicates. Both groups can be unambiguously identified and removed (56% of the initial dataset observations).

The second group of spam – spam or abusive tweets as defined by Twitter consist of so called non-exact duplicates (similar tweets created to share some link) and other spam tweets. We wrote a program that can identify and remove non-exact duplicates with the help of regular expressions (5% of the initial dataset).

Also we use information about account being suspended by Twitter to identify other tweets belonging to this group. Using these labels we tested several machine learning algorithms to predict probability of tweet being a spam and found that best results is provided by the logistic regression. Its final specification has 70% accuracy and 62% recall on a test set.

To identify messages belonging to the third group we proposed two ad-hoc criteria (using number of tweets in the dataset and the popularity of the account) and also build a neural network for prediction. After accounting for the class imbalance problem and testing for the sensitivity of results, our model provided 89% accuracy, precision and recall.

Although there was a lot that was achieved in the project, there is still room for an improvement. In particular, a larger dataset of tweets can be used. With the increase of the dataset size it can be expected that such algorithms as, for instance, neural network will outperform the logistic regression that currently provides the best results.

Another way to improve results is to use models to estimate values of the missing observations such as Account type.

Lastly the analysis can be further diversified by building a model to detect tweets produced by bots as they become increasingly common. To label such tweets several publically available datasets of bot tweets can be used.

Overview of work carried

Oleksa Stepaniuk - merging and pre-processing files, creating methodology for data cleaning (section 2.1), coming up with the idea for labelling and creating a web-scrapper, writing a presentation and presenting it, writing sections 2 and 5 of the report.

Adi Karri – supervising web-scrapping work, formatting presentation, conducting logistic regression analysis, writing introduction and section 4 (Biases and Validity), putting the report together, editing it.

Mansoor Muneer Reehana – writing methodology section, performing analysis with multiple classifiers and writing corresponding text (first part of section 3.1).

Chia-Yen Chiang – creating a neural network for the section 3.2, performing analysis of its sensitivity and robustness. Writing the section 3.2.

References

Heymann, P., Koutrika, G. and Garcia-Molina, H. (2007). Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. IEEE Internet Computing, 11(6), pp.36-45.

Varol, O., Ferrara, E., A. Davis, C., Menczer, F. and Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. In: The International AAAI Conference on Web and Social Media (ICWSM). [online] Available at: https://arxiv.org/abs/1703.03107 [Accessed 6 Dec. 2018].

Watson, C., "The key moments from Mark Zuckerberg's testimony to Congress", *the Guardian*, 2019. [Online]. Available: https://www.theguardian.com/technology/2018/apr/11/mark-zuckerbergs-testimony-to-congress-the-key-moments. [Accessed: 14- Jan- 2019].

Weiss, G. (2013). Imbalanced Learning: Foundations, Algorithms, and Applications. Chapter 2. Foundations of Imbalanced Data, 1st ed. The Institute of Electrical and Electronics Engineers, Inc., pp. 13-41.

Cs.cmu.edu. (2018). Zinman and Donath, CEAS 2007 - ScribbleWiki: Analysis of Social Media. [online] Available at: https://www.cs.cmu.edu/~wcohen/10802/fixed/Zinman_and_Donath%2C_CEAS_2007.html [Accessed 6 Dec. 2018].