University of Sussex

Computing for business and management

BSc

Trolls and controls: A temporal-based analysis of social media community clustering

Author
Henry Adams 132145

Supervisor
Dr. Luc Berthouze

07/05/2019

1 Declaration

'This report is submitted as part requirement for the degree of computing for business and management at the University of Sussex. It is the product of my own labour except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged.'

2 Abstract

Internet trolls have achieved much greater recognition within recent years than ever before [17]. Their impact on political and social discourse is a key area of research and discussion [4]. Many attempts have been made to study their effect on society [5], however for the purposes of this dissertation we will focus on ways in which they can be discovered [14]. Many approaches have been attempted to flag users who may be trolls however the majority rely on content based systems [9]. We have used a purely temporal approach to attempt to detect sub-communities within the social media site twitter.

Our analysis shows that it is possible to tag troll users with a purely temporal approach. However it is at a cost of accuracy. We can however measure this accuracy against the massive increase in efficiency this model achieves compared to content based models in order to elicit whether this model is novel.

Overall through community clustering [8] we are able to identify the features of clusters that are very likely to be trolls. We therefore focused on minimizing type 1 errors or false positives and discounting type 2 errors in order to ensure we only capture definite trolls and in some cases miss the rest. We can then use this information to remove definite trolls from the dataset and allow more sophisticated models to discern between the more complex cases. Overall our approach was able to capture around 90% of trolls in the dataset.

3 Acknowledgements

To Dr. Luc Berthouze for supporting and challenging me throughout the project.

Also to Joe Mountford, the entirety of 114 Ditchling road, and Propellernet, for keeping me sane.

Contents

4 Introduction

In the last 15 years social media has exploded in popularity with roughly two thirds of Americans reporting that they use Facebook. This figure is even higher for younger demographics [16]. Twitter is a large social media and discussion site. Everyday millions of tweets are created and viewed online [11]. With such a wide reach the site is used for many different purposes, however one more prominent use is the discussion of political news and events. With the space being so widespread it is able to directly reach and influence people around the globe [13].

With social medias popularity many news agencies are turning to the internet to post their content [12]. With the ability to reply to or comment on content from news agency users are able to discuss related content. This area for discussion has given many the opportunity to try to influence or change others views on political and social matters [21]. Often this is done by individuals but it is possible for interested parties to invade the twitter-sphere and maliciously influence others to better suit their needs.

This was most markedly demonstrated in the US political space with the term fake news being used often. Interested parties could manufacture content that may or may not be incorrect and is likely to influence readers political and social opinions [2]. This opinion manipulation can easily be seen as a political weapon, just as misinformation and propaganda were weapons previously [20]. Social media has much greater penetration in younger groups, many of whom are only just beginning to form political opinions and be given the ability to vote. There is therefore a need to deal with this type of tactic because it may be having an effect on people's political decisions and by proxy their lives [22].

The main question this raises is whether there is an efficient way to identify content created to manipulate opinion? And when this is possible what should be done with it? The latter question is more of a philosophical question, therefore i will be focusing on the more empirical first question. There have been many attempts with several different approaches. Such as Mihalyov, who used the number of times someone was called a troll to train a machine learning system to detect them [15].

This approach proved successful with an accuracy of 81-82 percent. There is also a more statistical approach that showed bots reposting hashtags had an effect on distributing fake news and propaganda [10]. However there are still limitations, For the purposes of this project I will be attempting to find a more efficient way to classify malicious propaganda accounts for social media platforms, specifically twitter, however this technique could be applied to other platforms.

5 Ethical Review

Due to the nature of my project and that it uses publicly available datasets with prior obtained consent it is unlikely that my methodology will contradict any points of the ethical guidelines, I will however explore point by point whether there is any oversight:

One concern is that the subjects may be exposed to risk greater than those in their normal working life. My study will only use pre collected data, it is therefore implausible that any user would be exposed to any issues due to my anonymous analysis of their data.

Participants data must also be collected on standardised hardware. Twitter is generally used on either PC's, mobile phones, or tablets. It is therefore implausible that my future use of pre-collected data is going to change this.

There is an issue of whether the user consents for the data to be used. When signing up to twitter users are asked to agree that their data is made public and will may be collected and stored, therefore their data is free to collate and analyse [19].

We must also ensure that the results may be used beyond the term of the project. Again within the twitter terms of service the user must specify that they accept their tweets are treated as public domain. They may however choose to delete the tweet so it is no longer viewable. I will therefore anonymise the users at a time when their tweets are still viewable to ensure that they cannot be traced back to the original user. We are also purely focused on the time at which the tweets were posted and will therefore discard all content when they are made anonymous

No incentives can be offered to participants. For the purposes of my project no incentives were offered; I gathered public domain data from users that I have no contact with.

We must not withhold information about the evaluation or materials intentionally. The creators of the data are a random selection from across the US, they were not informed of the study because their data was already public domain. No deception is intentionally used however it would be impossible to inform all users.

This study must not include participants under the age of 18. If a user is under 18 their date of birth will be hidden, I will therefore attempt to use only tweets from users who have a visible age meaning that they are over 18.

We must not include users with disabilities or impairments in our study. This information is not required to sign up to Twitter and would be impossible

to measure; likely being another study within itself. It is therefore beyond the scope of my investigation. This study must not include participants who myself or my supervisor have influence over. I will have no contact with the creators of the twitter data I have collected and therefore have no capacity to influence or pressure them.

All participants must be informed that they can withdraw at any time. A user may delete a tweet at any time they want, this may occur after the collection of my data, I will therefore immediately anonymise my data to ensure that it cannot be traced back to a user if they choose to revoke consent at a later date by deleting their tweet or their account. I am also only concerned with temporal data and will discard the content of the tweets. There will therefore be negligible data stored.

All participants must be informed of my contact details and my supervisor's. All the information is considered public domain and users have agreed in the twitter that their data can be used without any further consent. It would also be impractical for me to share my contact details with thousands of twitter users.

The evaluation must be described in detail with all participants. However for this study all the information is considered public domain and users have agreed in the twitter that their data can be used without any further consent Participants must be provided with sufficient information prior to starting the session. The creators of the data gave consent for their data to be used without any further notice or consent.

All data must be stored securely in an anonymous form. Therefore all data I have and will collect from twitter will be anonymised into a time series with a unique identity number I generate, a tag with which I can denote if they are a troll, and the actual twitter behaviour. This will be stored on my personal computer which requires a fingerprint to access and is equipped with industry standard anti-virus software. I will not perform analysis on any untrusted networks and complete work within the university's secure network or my own secure network.

6 Project Summary and Specifications

With the potential for internet trolls to effect opinion online in an organised and malicious way it can be argued that we should be able to detect these trolls in an efficient manner. Whilst others have attempted to solve this issue through content based approaches [9], my working hypothesis is that by analysing the temporal behaviour with which twitter communities tweet in response to events, It may be possible to ascertain whether trolls can be identified with a good degree of accuracy. With context we are able to identify whether accounts are likely bots or organised groups. By using a temporal approach we are able to strip away the majority of the data set and greatly alleviate any problems related to processing times and efficiency.

The majority of research in this field focuses upon the content of propaganda rather than the behaviour of propaganda creators and distributors [9], I will focus on the latter approach to try to discover whether it could be more efficient. Community structure detection, the act of grouping together individuals dependent on their behaviour, is widely used in machine learning and network science [8]. It has recently been applied to the brexit debate and pro-leave bots [10]. And has also been used for commercial purposes to identify their demographic for the purposes of marketing [18]. Recently the vast amount of political discourse in the world has led to the revelation that twitter and other platforms have been used for psychological manipulation [3]. It is therefore important to be able to identify such manipulators in an efficient manner, by focusing on behaviour rather than content it may be more possible to do so. In this project I will seek answers to research questions such as:

- Can you identify genuine twitter accounts purely from their behaviour?

  By using frequency analysis and clustering are we able to discern two distinct communities of propaganda and non propaganda accounts.

- Would these methods be more efficient than analysing content?

  Comparison of more content based classification systems that use clustering and whether time frequency methods alone are faster.

- If they are more efficient is their time benefit worth their potential loss in accuracy?

  Comparison of the two methods. Whilst the frequency system is likely more efficient is there any loss of accuracy and is this loss of accuracy significant.

I aim to use data gathered from known troll accounts and data from regular users, I will then use known clustering models to see if they are able to identify users by time series alone. By using correlation and dependence we may be able to cluster twitter users based on their temporal behaviour. We can then see if these groups correlate with known disingenuous accounts. It may also be

possible that these accounts will work to act as genuine accounts, in such a case depending on their success we would expect two extreme clusters or little to no differentiation from regular data.

7 Literature review

7.1 Current Goals

My current goal is to find out whether you can classify a twitter user as a troll purely by the temporal behaviour with which they tweet. I will therefore attempt to discover communities within the larger collection and see if we can identify a community solely made up of trolls.

7.2 Current Approaches

There are several ways in which this has been attempted before. The main analysis I have encountered has been a content based approach, where the tweets would be downloaded and a model would be trained and built to detect features such as sentiment and opinion. This proves to be accurate in circumstances where there has been sufficient time to develop a complex model. However the cost of developing the model is very high in all cases [7]. The approach of purely temporal data has been expansively covered in many areas such as geology and other areas of data science, [6] However there is little to no instances of this approach being applied to twitter data. Many complex machine learning techniques trade of accuracy for efficiency, therefore being able to train a quick and fairly accurate model should be deemed novel, as it would allow us to apply complex analysis to extremely large datasets without encountering processing problems as frequently.

There have already been attempts to streamline the way in which twitter content is analysed. One tested approach is measuring an accounts likeness to be a troll predicated upon other users reaction to them [15]. researchers used this method to analyse whether someone was a troll based on how many times in a thread they were called a troll by others. They were only concerned with word frequency within replies and could therefore cut the processing time massively. The experiment conducted upon users within a Bulgarian news forum showed that they could train a machine learning model based upon this data with a good degree of accuracy (82-95) by using a rich feature set.

This method allowed them to bypass several hurdles such as cultural language and idioms within a language that machine learning may be unable to accurately classify. The use of user perception allows them to allocate simple metrics to each account such as word frequency giving a reasonable level of accuracy and allows them decide whether they are likely to be a troll. This gives them an easy to understand dataset however it does have shortcomings. For example it would be easy to reply to an accusation of being a troll with a retort accusation of the non-troll being a troll. False accusations could quickly muddy the waters of accuracy within this method. There is also an aspect of group mentality, where emotions are charged people may react aggressively online and garner the support of other, by grouping together they can quickly

change whether someone is likely to be a troll. It may also be argued that if someone is called a troll enough they may change their behaviour, if someone is falsely accused of trolling enough could this cause them to react in troll like fashion? This research was however conducted in an area where it is likely that Russians may try to influence. When there is a prevalence of trolls they may be easier to spot.

We may also apply a statistical approach to the problem. Howard investigates whether by monitoring hash-tags with potential for troll use; they can in turn identify the trolls themselves. In the time leading up to the Brexit debate twitter was a space for political news and discussion. By monitoring the usage of key hash-tags they were able to identify bots and their impact [10]. It is made clear that the average users is unlikely to be able to tell the difference between a bot and a regular user. With the bots at their disposal smaller groups may be able to dominate conversation with twitter and ultimately sway opinion. There is however an issue that it is difficult to track whether hash-tags are being used by bots or normal users, it may however be possible to explore whether re-tweet or original content plays into the spread of hash-tags.

A simpler content machine learning approach may also be possible. By analysing the way in which trolls craft their tweets in terms of content we can train a machine learning model to detect trolls with some degree of access. This is approach is widely used in order to mine opinion and sentiment on twitter. This is a excellent approach for finding sentiment on a topic and could easily detect whether a user is right wing or left wing based on their opinions, it does however struggle to define a fake right wing account to spread discord and a genuine right wing person. This is because it only looks at the content on a scale of right to left wing. We could however train it to look for content based on its credibility, this would help us to find occurences of fake news used to spread discord. The main issue with this approach is that trolls could rely on credible sources and sway them with logical fallacies in order to create discord.

I would therefore look to pursue a temporal based approach based upon my hypothesis. By using temporal clustering methods we can hope to find a community of trolls. This approach has been applied in a different way by using clustering of temporal twitter data to detect the occurrence of events. We can therefore work with this approach from known events and apply some of the methodology in an opposite direction. Instead of detect events we can mark how different users react to events and whether one community can be identified as malicious by using tagged data.

7.3 Conclusion

Through analysis of the previous papers I can conclude that this approach is novel in that this temporal based analysis has not been applied to twitter be-

haviour previously. I am able to take methods from other papers to improve my methodology and analysis however I do have scope to make this project unique. The main issue I have seen with the previous methods is that the are not directly suited to my application or that if they would be effective they would be inefficient and difficult to use.

Good machine learning systems such as those used by Agarwal [1] analysis are excellent at determining sentiment and data mining, but lack the efficiency to detect malicious intent on a large scale. They would prove to be more accurate than my system however the time spent to train the model makes them difficult to implement. In contrast others such as Howard explore a small subset of bots based upon a their use of hash-tags. This analysis is fruitful but is still content based and would have to involve training a complex model which is far less efficient than a purely temporal approach. It is unlikely for me to be able to achieve accuracy levels as good as or better than the content based systems, however, the trade-off increase in efficient may prove that this method is useful.

The most informative approach I have found was the temporal data used to identify key events used by Becker, by applying this methodology in a different direction we can work backwards from known events and instead apply the clustering to the users reactions in order to better create homogeneous communities of which hopefully one or more will be solely trolls. Overall I am able to continue developing my hypothesis and use more ideas from papers and journals, and create a more informed hypothesis.

8 Methodology

8.1 Overview

Overall we will use the temporal data of a subset of twitter users. We will then make the data anonymous and strip away all content except for the temporal data. We do this in order to try and build an extremely limited data set which will allow us to process it more efficiently. We will attempt to build a feature set based upon their temporal behaviour. Including behaviours such as seasonality and periodicity. Once these features are defined we are able to apply a hierarchical clustering method in order to group users with similar behaviour, and find a subset purely made up of trolls. We must then compare the accuracy we have achieved with standard content based approaches and then compare our efficiency.

8.2 Data Collection

For the purposes of this project we will use the web-based tool Brandwatch (Audiences) in order to generate audiences to generate a list of followers from each candidate which we can then collect a sample from. We may then use another component of Brandwatch (Insights) to collect the time lines of tweets of the elected users from which we may determine their temporal behaviour. This methodology was applied to key figures of the 2016 US election:

Democrat	Republican
Hillary Clinton	*Donald Trump*
Bernie Sanders	*Ted Cruz*
Martin O'Malley	*John Kasich*
Lincoln Chafee	*Marco Rubio*

We have selected 4 candidates from each party who had the highest percentage of the popular vote and will use their followers. We chose these users because they have shown at least some interest in politics by following one or more of the presidential candidates and can therefore be used as a control group. There will likely be an internal spectrum of people with only a small interest all the way up to those with extreme interest.

The key issue which is immediately apparent is that they have vastly different follower counts, from less than a million, up to sixty million. Trump has more followers than all the others combined. we will therefore take an equal distribution across all candidates to ensure we have a balanced demographic to begin with.

We will use this control data in comparison to a tagged list of known troll accounts compiled by twitter for the House Intelligence Committee's investigation into Russian interference in the 2016 election.**REF**

8.3 Data Pre-processing

The Pre-processing step will be used to strip down our data and to make it anonymous. We will remove all data except for:

- A unique integer identifier for each user
- A list of date times with which each of their tweets was posted

```
user_id
18710816.0    2016-11-04 14:01:41
18710816.0    2016-09-14 04:06:46
18710816.0    2016-10-03 18:46:57
18710816.0    2016-10-23 18:33:13
18710816.0    2016-07-26 07:17:12
Name: created_str, dtype: object
```

We are then able to transpose this from a list of timestamps associated with users to a table with users time lines grouped.

time	2014-07-17 20:27:52	2014-07-22 08:46:37	2014-07-22 17:45:07	2014-07-22 18:26:47	2014-07-28 15:05:01	2015-01-31 09:53:48	2015-01-31 09:54:02	2015-01-31 09:54:17	2015-02-20 12:19:49	2015-02-20 12:20:50	...
user_key											
austinlovesbeer	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...
dorothiebell	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
finley1589	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
giselleevns	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
hollydler	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

Figure 1: Transposed user data

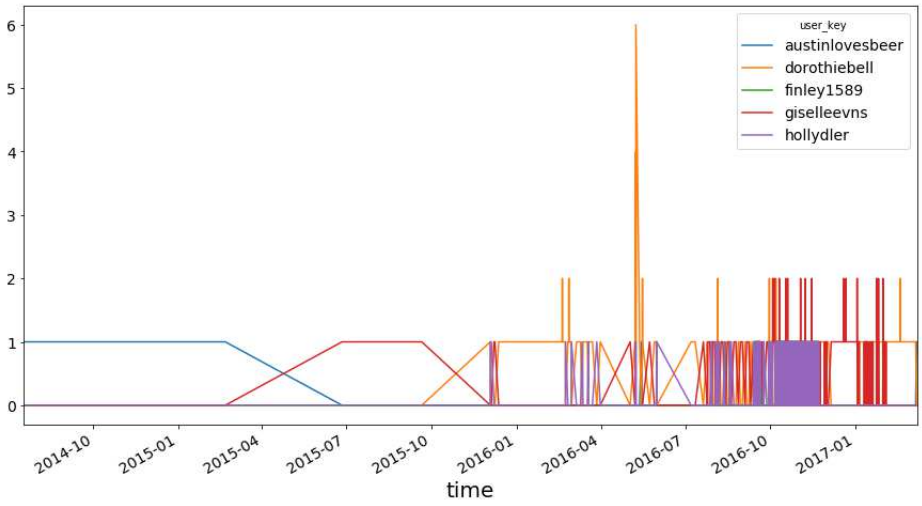


Figure 2: Users as time series

we can then develop a frequency time series graph for each user, to measure their temporal behaviour. From each time series of each user why apply an autocorrelation function I order to measure the variance over time and try to pick out any elements of seasonality. This will create a numPy array which we can then further process.

8.4 Feature Set

We aim to build a feature set that is able to identify key behaviours which are associated with trolls. We have defined these features as high variance within their tweeting pattern, for example tweeting 50 times in an hour then nothing for several days. We can map this using the autocorrelation function.

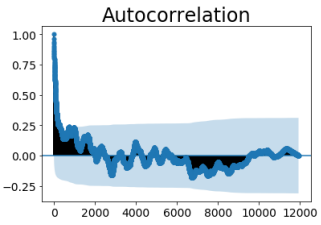


Figure 3: High variance autocorrelation

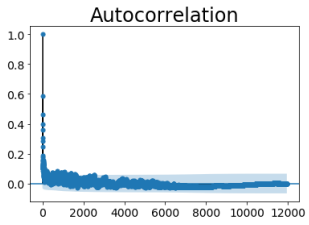


Figure 4: Low variance autocorrelation

In the left example we are able to see a time series with high variance, whereas in the right we can identify a much lower variance. This is important because this type of behaviour is not human-like, mostly users will have a steady amount of tweets throughout their time-line. There may be some variance, but not manifested in such an erratic way as we see here.

We must also analyse their modal behaviour based upon time windows identifying periodicity.

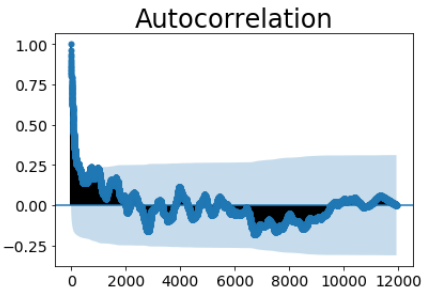


Figure 5: High variance seasonality

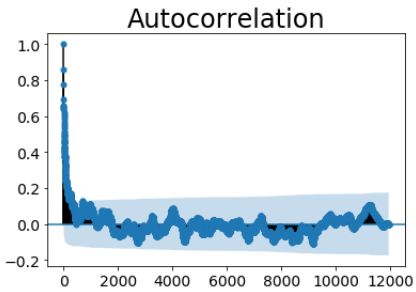


Figure 6: Low variance seasonality

In the left example we can see similar spikes and dips with the right example, albeit with a lower variance. From these similarities we are able to measure a high degree of similar periodicity. i.e. The users commonly tweeted with similar gaps of a period of time, e.g. every three days or once a week, but not necessarily on the same day.

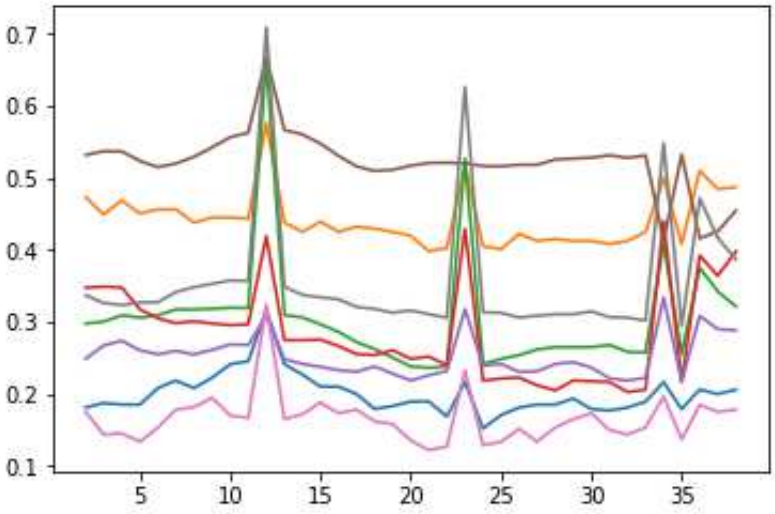


Figure 7: event based seasonality

In the fig 7 we can see a clear exhibition of seasonality within our autocorrelation. This is the main feature we will attempt to find within our clusters after we have located those with high variance.

We search for time series' where we see high variance exhibited, for most normal tweeters we would expect them to tweet a fairly consistent number of times per day or per week. This may increase due to external circumstance but we generally shouldn't see too much deviation from the average. We therefore must look for instances where accounts tweeted exponentially higher in one period then drop off completely. Using autoregressive models we are able to map these instances with higher frequencies representing higher variance. We can then find clusters that exhibit high variance but all have similar seasonality.

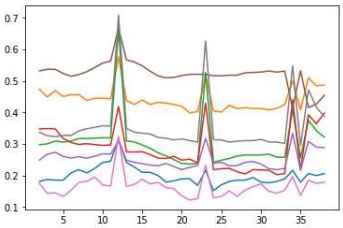


Figure 8: seasonal high variance

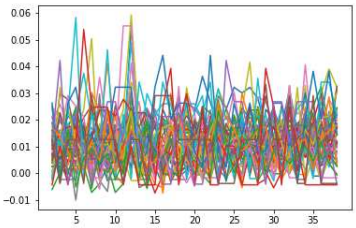
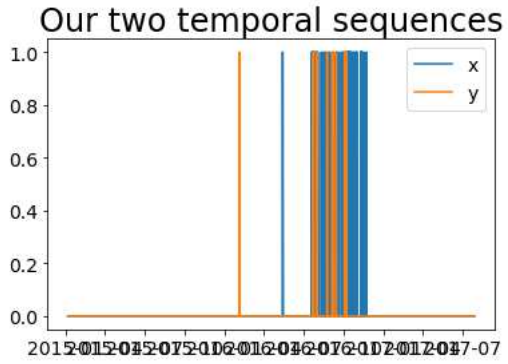


Figure 9: messy high variance

We could then apply this methodology further with different time windows applied, for instance is there a large burst of tweets every day, every week, or every month, however for the purposes of this study we will use one time window, to ensure that the processing time does not become too large. By comparing behaviours we are able to more accurately compare behaviour of tweeters. We must also assume that there will be some seasonality present in our data set. For example many tweeters may have jobs, so we would expect there to be fewer tweets within working hours except for breaks. We will therefore be able to group users of similar circumstance.

8.5 Clustering

Applying autocorrelation will give us a graph for each feature, we can then compare these to find the most similar. We must first define our two time series and plot them.



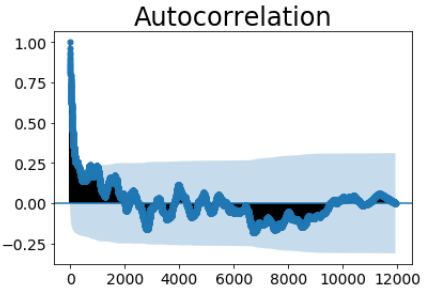


Figure 10: Autocorrelation of x

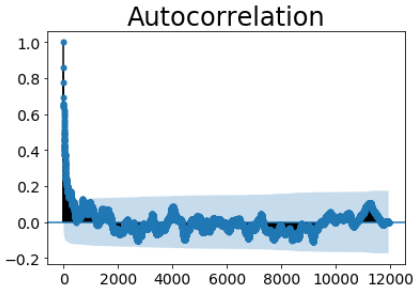


Figure 11: Autocorrelation of y

we must then define the autocorrelation function for each time series as shown in fig 10 and 11.

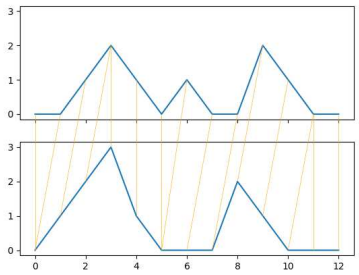


Figure 12: dynamic time warping comparison

Once we have achieved this we must compare the two using dynamic time warping. Dynamic time warping is able to measure the similarity between two temporal sequences even if they vary in speed.**??**. We find the two most similar and group them.

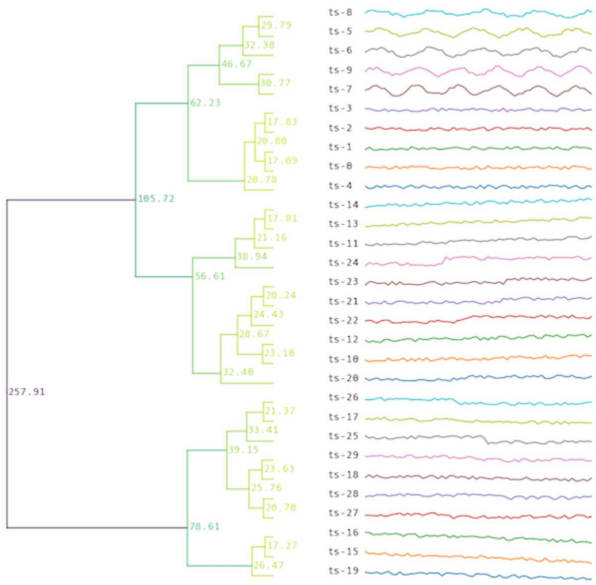
we must then concatenate our series for each autocorrelation function into a matrix:

```
[[ 1.00000000e+00 -1.90512475e-04 -1.90524570e-04 ... -1.90959972e-04
  -1.90972066e-04 -1.90984161e-04]
 [ 1.00000000e+00  1.39798618e-01  5.01941039e-02 ...  5.01859490e-02
   3.22648650e-02  3.22646385e-02]
 [ 1.00000000e+00  6.83823767e-02  1.01560133e-02 ...  2.47024779e-02
   1.01456776e-02 -4.41112275e-03]
 ...
 [ 1.00000000e+00  4.63331585e-01  2.84442038e-01 ... -1.78512696e-03
   3.39926468e-02  6.97704205e-02]
 [ 1.00000000e+00  8.07498455e-01  7.09238784e-01 ...  3.05934224e-01
   3.01916427e-01  2.93888338e-01]
 [ 1.00000000e+00  1.69354870e-01  1.36087173e-02 ...  1.36002700e-02
  -3.70506659e-03  1.35998008e-02]]
```

This will show the output of the autocorrelation function for each series which we can then use to cluster by then iterating this method over the entire matrix to find the most similar time series and cluster them together. we are then able to hierarchically cluster them using dynamic time warping to compare each graph and find its most similar neighbour which we we can then group.

8.6 Cutting the Dendogram

Once we have achieved a suitable dendogram we must seperate it into obvious clusters. We must therefore look for groups with behaviour we expect of trolls. In the below example the top cluster stands out with ts[8,5,6,9,7,3]



more importantly than just high variance is a high degree of seasonality, we have found during this study that this is more indicative of troll behaviour than just high variance alone. We can also see a degree of seasonality in our sample ts[8,5,6,9,7,3]. This type of bahaviour is likely to map to key events within the data, In this case likely mapping to key political rallies and votes.

8.7 Logistical regression

we must then apply logistic regression to the model [**?**]. For the purposes of this project we will use pre-built modules within the python module scikit learn [**?**]. We must first create a test train split within the data so that we can test on data the model has never seen to ensure it is fair. From applying this logistic regression through the use of the time series and its clustering list we are able to use the clusters marked as trolls to train the model into fairly accurately predicting trolls.

9 Results

over all using our system we were able to successfully generate several clusters. With tuning we were able to cut the dendrogram at a depth that gave us meaningful data. Some clusters were mixed with both trolls and controls, the majority were controls purely and a select few were purely trolls.

cluster	control	troll
15	8	13
1	0	11
16	0	8
13	53	3
2	8	2
4	2	2
5	0	2
3	4	1
6	5	0
7	20	0
8	6	0
9	11	0
10	1	0
11	1	0
12	52	0
14	1	0

cluster	control	troll
13	53	3
12	52	0
7	20	0
9	11	0
2	8	2
15	8	13
8	6	0
6	5	0
3	4	1
4	2	2
10	1	0
11	1	0
14	1	0
1	0	11
5	0	2
16	0	8

9.1 Accuracy

9.1.1 type 1

We will almost certainly find a diminished level of accuracy in comparison to a text analysis based model. We will therefore focus on minimizing type 1 errors and focus on the behaviour of definite trolls, whilst leaving possible trolls for more sophisticated models. by using this methodology we are able to identify clusters with clear seasonality. In this case clusters 1 and 16 stand out as having the highest percentage of true trolls. by analysing their graphs we are able to see that they worked concurrently around key events. This alludes to the hypothesis that there are agents working concurrently to influence social media through propaganda.

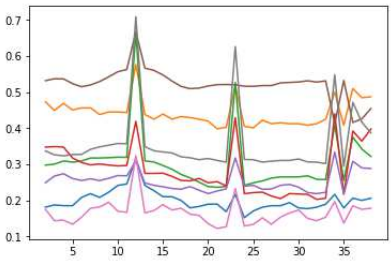


Figure 13: cluster 16

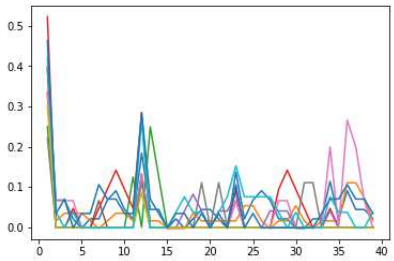


Figure 14: cluster 1

By looking at these two clusters we can clearly see a high degree of seasonality within their variance. We can see three key events in fig 13 which will likely coincide with major political events. In figure 14 the results are less obvious but still show 2 key events which coincide with the events in figure 13. we can then compare these to clusters that had little to no trolls in them.

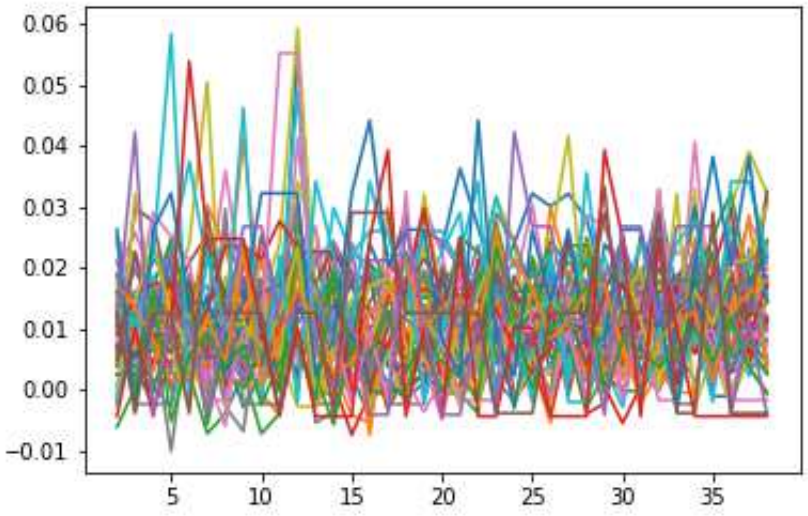


Figure 15: cluster 12 - no trolls

as we can see in figure 15 there are no real key events that vary too greatly from other spikes.

9.1.2 type 2

For the purposes of this study we will likely not focus on minimizing type 2 errors. It would be useful to gain high levels of accuracy across all levels, however due to limitations of the sophistication of this model it is more suited to capturing the more obvious trolls and leaving more complex cases to a more complex model.

9.1.3 Issues with accuracy

Overall our clusters can be separated into three groups:

- No real seasonality
- Potential seasonality without concurrency
- Extreme seasonality with concurrency

By focusing on clusters that fit our hypothesis we are able to gain high degrees of accuracy. Overall we are able to quickly tag 50% of trolls. Whereas the other 50% of our trolls are spread across many clusters with other users.

cluster	control	troll
13	53	3
2	8	2
15	8	13
3	4	1
4	2	2

9.1.4 Overall Accuracy

Overall after logistical regression we are able to acheive the confusion matrix of:

predicting on test data

Actual \ Predict	0	1
0	13	2
1	1	12

we are also able to achieve an overall accuracy of:

Overall ACC 0.89286

Whilst we are not able to accurately tag a high degree of trolls we are able to use this methodology as a good initial starting point to then expand upon and use other models. while this seems like a fairly low accuracy score for a machine learning system it must be taken with a pinch of salt, as this is exponentially faster than content based models to train and implement.

10 Discussion

10.1 Findings

overall we agree with previous research, however this method suits a different purpose. It is useful in that it is able to quickly tag data within a large dataset with passable accuracy. Which can then be passed on further along a pipeline to more in depth analysis of the content.

10.2 Goals achieved

we were able to achieve a good degree of accuracy for the amount of time the model would take to train. However in the world of machine learning the accuracy level would be seen as sub-par, it is rather a best case scenario when content based approaches would take too long to train.

10.3 Further work

To further increase the accuracy of the model a more diverse feature set could be implemented, this model relies on a very small feature set to increase efficiency however more work could be done to find the sweet spot between efficiency and accuracy.

11 Conclusion

Overall this project was able to cluster communities within a dataset and identify the features of these clusters that were considered troll-like. We were able to achieve a good degree of accuracy in comparison to other machine learning content based approaches. We did notice a slight drop in accuracy however this makes up for the extreme gain in efficiency.

I believe this project shows a novel approach to troll detection, making using of community clustering in a way it has previously not been used.

It is limited by its accuracy by using a more concise dataset, however this can be expanded with a greater feature set. This project only used a small feature set but expanding this would likely lead to greater accuracy with little increase in time taken.

References

[1] Basant Agarwal and Namita Mittal. Machine learning approach for sentiment analysis. In *Prominent feature extraction for sentiment analysis*, pages 21–45. Springer, 2016.

[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[3] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265. IEEE, 2018.

[4] Jonathan Bishop. Representations of trolls in mass media communication: a review of media-texts and moral panics relating to internet trolling. *International Journal of Web Based Communities*, 10(1):7–24, 2014.

[5] Matin Chiregi and Nima Jafari Navimipour. A new method for trust and reputation evaluation in the cloud environments using the recommendations of opinion leaders' entities and removing the effect of troll entities. *Computers in Human Behavior*, 60:280–292, 2016.

[6] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. VS-PETS Beijing, China, 2005.

[7] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45. ACM, 2003.

[8] Leonardo N Ferreira and Liang Zhao. Time series clustering via community detection in networks. *Information Sciences*, 326:227–242, 2016.

[9] Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

[10] Philip N Howard and Bence Kollanyi. Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum. *Available at SSRN 2798311*, 2016.

[11] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[12] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[13] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.

[14] Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 443–450, 2015.

[15] Todor Mihaylov and Preslav Nakov. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 399–405, 2016.

[16] Andrew Perrin. Social media usage: 2005-2015. *Pew Internet and American Life Project*, 2015.

[17] Pnina Shachaf and Noriko Hara. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370, 2010.

[18] Andrew N Smith, Eileen Fischer, and Chen Yongjian. How does brand-related user-generated content differ across youtube, facebook, and twitter? *Journal of interactive marketing*, 26(2):102–113, 2012.

[19] twitter LLC. twitter terms, 2019.

[20] Tobias Ursprung. The use and effect of political propaganda in democracies. *Public choice*, 78(3-4):259–282, 1994.

[21] Maurice Vergeer, Liesbeth Hermans, and Steven Sams. Online social networks and micro-blogging in political campaigning: The exploration of a new campaign tool and a new campaign style. *Party politics*, 19(3):477–501, 2013.

[22] Miranda Yates and James Youniss. Community service and political identity development in adolescence. *Journal of Social issues*, 54(3):495–512, 1998.

12 Appendices

12.1 Accuracy on Testing Data

predicting on test data

Predict	0	1
Actual		
0	13	2
1	1	12

Overall Statistics :

95% CI	(0.77829,1.00742)
AUNP	0.89487
AUNU	0.89487
Bennett S	0.78571
CBA	0.8619
Chi-Squared	17.37436
Chi-Squared DF	1
Conditional Entropy	0.48514
Cramer V	0.78773
Cross Entropy	1.0
Gwet AC1	0.78599
Hamming Loss	0.10714
Joint Entropy	1.48145
KL Divergence	0.00368

Kappa	0.78571
Kappa 95% CI	(0.55659,1.01484)
Kappa No Prevalence	0.78571
Kappa Standard Error	0.1169
Kappa Unbiased	0.78544
Lambda A	0.76923
Lambda B	0.78571
Mutual Information	0.51486
NIR	0.53571
Overall ACC	0.89286
Overall CEN	0.44355
Overall J	(1.6125,0.80625)
Overall MCC	0.78773
Overall MCEN	0.35215
Overall RACC	0.5
Overall RACCU	0.50064
P-Value	6e-05
PPV Macro	0.89286
PPV Micro	0.89286
Phi-Squared	0.62051
RCI	0.51677
RR	14.0
Reference Entropy	0.99632
Response Entropy	1.0
SOA1(Landis & Koch)	Substantial
SOA2(Fleiss)	Excellent
SOA3(Altman)	Good
SOA4(Cicchetti)	Excellent

Scott PI	0.78544	
Standard Error	0.05845	
TPR Macro	0.89487	
TPR Micro	0.89286	
Zero-one Loss	3	

Class Statistics :

Classes	0	1
ACC(Accuracy)	0.89286	0.89286
AUC(Area under the roc curve)	0.89487	0.89487
AUCI(Auc value interpretation)	Very Good	Very Good
BM(Informedness or bookmaker informedness)	0.78974	0.78974
CEN(Confusion entropy)	0.43358	0.45425
DOR(Diagnostic odds ratio)	78.0	78.0
DP(Discriminant power)	1.04317	1.04317
DPI(Discriminant power interpretation)	Limited	Limited
ERR(Error rate)	0.10714	0.10714
F0.5(F0.5 score)	0.91549	0.86957
F1(F1 score - harmonic mean of precision and sensitivity)	0.89655	0.88889
F2(F2 score)	0.87838	0.90909
FDR(False discovery rate)	0.07143	0.14286
FN(False negative/miss/type 2 error)	2	1
FNR(Miss rate or false negative rate)	0.13333	0.07692
FOR(False omission rate)	0.14286	0.07143
FP(False positive/type 1 error/false alarm)	1	2
FPR(Fall-out or false positive rate)	0.07692	0.13333
G(G-measure geometric mean of precision and sensitivity)	0.89709	0.8895

GI(Gini index)	0.78974	0.78974
IS(Information score)	0.79355	0.88452
J(Jaccard index)	0.8125	0.8
LS(Lift score)	1.73333	1.84615
MCC(Matthews correlation coefficient)	0.78773	0.78773
MCEN(Modified confusion entropy)	0.625	0.64804
MK(Markedness)	0.78571	0.78571
N(Condition negative)	13	15
NLR(Negative likelihood ratio)	0.14444	0.08876
NPV(Negative predictive value)	0.85714	0.92857
P(Condition positive or support)	15	13
PLR(Positive likelihood ratio)	11.26667	6.92308
PLRI(Positive likelihood ratio interpretation)	Good	Fair
POP(Population)	28	28
PPV(Precision or positive predictive value)	0.92857	0.85714
PRE(Prevalence)	0.53571	0.46429
RACC(Random accuracy)	0.26786	0.23214
RACCU(Random accuracy unbiased)	0.26818	0.23246
TN(True negative/correct rejection)	12	13
TNR(Specificity or true negative rate)	0.92308	0.86667
TON(Test outcome negative)	14	14
TOP(Test outcome positive)	14	14
TP(True positive/hit)	13	12
TPR(Sensitivity, recall, hit rate, or true positive rate)	0.86667	0.92308
Y(Youden index)	0.78974	0.78974
dInd(Distance index)	0.15393	0.15393
sInd(Similarity index)	0.89115	0.89115