Sentiments Spreading over Twitter

How do the sentiments of an user affect another user?

Hamdan Ahmed
ahmed@uni-bremen.de
Universität Bremen

Pedro Dias
pmartins@uni-bremen.de
Universität Bremen

Pedro Roque
pedro1@uni-bremen.de
Universität Bremen

Ricardo Martins
ricardo1@uni-bremen.de
Universität Bremen

ABSTRACT

This study explores the propagation of sentiments between users on Twitter. The hypothesis is that sentiments are propagated between users when one user writes a tweet that demonstrates a certain sentiment and other users are persuaded to feel and demonstrate the same sentiment by reading the tweet. To test the hypothesis tweets wrote on a part of London were collected and analyzed. A network of the users that wrote the tweets was created and a score was assigned to each user based on the sentiment that the user expressed on the tweets that he wrote. Communities were then found on this network and each community was then analyzed to check if the majority of the users present in the community shared the same sentiment which would be an evidence that there is sentiment propagation. At the end, the results were a good proof that there is sentiment propagation since the majority of the communities found share the same sentiment but this proofs can also be arguable since the sentiment analysis of a text has limitations and we can never be 100% certain that one user's sentiment was influenced by reading the tweet from another user. Nevertheless, we have a strong conviction that based on our results the hypothesis of this study is true.

1 INTRODUCTION

Human interactions have been present in the human species since it existed. Nowadays there is proofs that even since the first time of the human species, humans gathered and did activities together. So as far as it's known, humans always had social interactions between them for different reasons such as interactions with the purpose of survival or just simple friendship interactions. All this social interactions create social networks which are present in our day-to-day life making this networks an important subject to be analyzed and studied.

Nowadays, there is a lot of different social media that empower the creation of social networks on a bigger scale than decades ago. Some example of the most famous social media websites are Twitter, Facebook, Instagram, Tumblr, LinkedIn, Google+ and Youtube.

All of this social media are different but they all have one fact in common, they boost the creation of big social networks by making it easier for their users to interact with each other.

On this study, the interactions between the users of Twitter were analyzed. Twitter is one of the biggest social media present in our world where users communicate in short texts called 'tweets'. Each user has a profile that works as personal blog. In the user's profile the tweets of the user are shown by a chronological order being the ones on the top the most recent. This tweets are visible to every other user that visits the user's profile if the user has the profile as public, otherwise if the user as the profile as private, only the other users who follow this user can see his tweets.

On Twitter there is a basic link that is used to create a user's social network. This link is 'following'. When a user wants to see another user tweets he follows him. So because of this fact, each user has followers and follows people. Each user also has a 'feed' where the tweets from the other users who the user follows are shown also in a chronological way.

Users can interact with each other directly by mentioning other users or by 'retweeting' them. Retweeting is the act of sharing a tweet made by other user and publishing it on the user's (the one that retweeted the tweet) personal profile.

Twitter has a great potential to be used to study social networks because of the way it connects users and also because of the texts each user publishes. For this reasons this study is done based on Twitter.

Freedom of speech is one of the most important rights achieved in our society and Twitter is a platform where this freedom of speech is boosted since users can write whatever they want on their tweets. They can either write what they're feeling, write about a specific topic and give their opinion or just write to each other.

From this tweets we can find a lot about what is going on about a specific subject and even about the state of mind of the user who wrote the tweet.

Based on this type of information that can be found on tweets, this study aims to prove that the content of a tweet can influence the sentiments of an user that read the tweet. If the hypothesis by the end of the study is confirmed then, for example, happy tweets could be used to influence sentiments of another users with the purpose of influencing their consuming behaviour since it is proven that happiness and consumer behaviour are related[2].

Another possible outcome of this study could be that if the sentiments are proven to propagate then, for example, people with depression would have proof that following users on Twitter who normally are tweeting sad tweets is bad for their recuperation and they should avoid it.

This study tries to find the best proof it can that there is spreading of sentiments between users over the tweets that each user publish and tries to do this by firstly gathering tweets from a part of London and then building a network based on the data gathered.

The nodes of the network created represent the users and the edges between the nodes represent mentions/retweets between the users. Each user's tweets will be analyzed and a sentiment score will be assigned to each user. After this is done, community finding algorithms will find communities in the network and each of the communities will be analyzed to check if the majority of the users share the same sentiment. This will be done by two different algorithms of community finding to demonstrate that the results found are not influenced by the community finding algorithm chosen.

If in different datasets the majority of the clusters have the users inside it with the same feeling then we get a strong proof that sentiment is propagated through the tweets between users which makes the hypothesis this study aims to prove true.

At the end of the study we found out strong proofs that the hypothesis was right since in the majority of the clusters the users share the same sentiment as this report shows in the results chapter. Although the study finds strong proofs which indicates that the hypothesis is correct, it didn't find a 100 percent certain proof because this is topic with imprecisions and subjectivies that need to be taken in account.

2 BACKGROUND

In this study a tool for sentiment analysis made from other authors was used. Two community finding algorithms and a tool to help on the creation of the networks from other authors were also used. The tools and the algorithms to find communities are described in the next paragraphs of this chapter.

The tool used for sentiment analysis of each tweet was vaderSentiment[3].

VaderSentiment is a tool which is free to use and open source that is available on GitHub or even on pip where it can be installed if changes on the tool are not necessary. This tool classifies a text based on a dictionary with a big amount of words/symbols and where each word/symbol has a score representing the sentiment associated with the word/symbol. The tool takes in consideration word boosters, i.e., words that increase the value of the next work, for example if a text is 'very happy' then the happiness score attributed by the tool will be higher than the happiness score attributed to the text 'happy'. It has a really extensive dictionary and even includes emojis but it only has one language available, English.

At the end of analyzing a text the tool gets 3 values from the words that it read, positive (happy), negative (sad) and neutral. This values are in the scale of 0 to 1. From this values the tool then calculates a final value that represents the sentiment of the text on a scale from -1 to 1 being -1 the extreme of sadness, 1 the extreme of happiness and 0 the neutral.

The tool as it is on GitHub receives a text input in the terminal, classifies it and writes in the terminal the classification. Instead of this, it was necessary that the tool received a file with several texts to classify and to wrote the classifications in another file, so, due to this fact, the open source code was changed.

For the creation and manipulation of the networks created, NetworkX[1] was used which 'is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks'.

One of the algorithms used for finding communities in the network was the Louvain method for community detection[5]. This algorithm focuses on the optimization of the modularity as the algorithm progresses. Modularity is a value that represents the difference between the density of edges inside communities to edges outside communities and optimizing this value gives the best possible communities of a network. The algorithm follows a greedy approach which starts by finding small communities and then representing those communities found by a node and repeating the process until all communities possible are found. This algorithm is one of the fastest community finding algorithms available nowadays.

The other algorithm used for finding communities in the network was Girvan-Newman algorithm[4] which takes a different approach and detects communities by progressively removing edges from the original network. The connected components of the remaining network are the communities. This algorithm chooses the edge to remove based on the value of edge betweenness. Edge betweenness is defined as the number of the shortest paths that go through an edge in a graph or network[4]. Because of this, when choosing an edge to remove, the algorithm starts by removing the edge that has the highest value of edge betweenness because this edge is the one who is most likely to be between communities.

3 DATA EXTRACTION

Twitter API was used to get data that was used by the program built in order to analyze it and withdraw conclusions from it.

Twitter API has three ways of accessing the API: Standard, Premium and Enterprise. The one used to access Twitter API was the Standard way since it was the free one. And because it is the free one it has some limitations that changed the way this study was conducted.

The data got from Twitter API consists on a big file with thousands of tweets (the number of tweets to retrieve can be defined) and to get this tweets queries had to be used on the Twitter API. This queries can have a lot of parameters which act as filters in the tweets the query retrieves, for example geolocation (only tweets made in that geolocation are retrieved) or just a word (only tweets with the specific word in the text are retrieved). One of the available parameters can be used to retrieve tweets done by a specific user. Since this parameter was available and it could be used, the initial plan of the approach to prove the hypothesis was based on getting the data from Twitter API using this parameter.

Data would start by being retrieved based on a specific user and then the Twitter API would be used to get the list of users that the initial user followed. This (getting the users a specific user follows) is also possible and the Twitter API allows the standard access to do it. After this steps, the user list of the users that the initial user follows would be checked and each user in that list would be subject to a query to retrieve his tweets (same query that was used on the initial user but with the parameter now being this user). This would be done until the second level of neighbors (the people that the first user follows are his first neighbors and the persons who the people that the first user follows follow are the second neighbors).

If the data extraction was done like this, an edge of the network would be created between two users if two users followed each other.

The problem was that doing this with the standard access to the Twitter API didn't work as it was expected. Standard access to Twitter API has many limitations, there is a limit of queries that can be done per month and with the previously explained way of getting data there would be a need to execute more queries than the ones that are allowed. Another limitation is also the fact that standard access only allows to get tweets made in the last seven days and it doesn't guarantee reliable data, i.e., imagine a user wrote 50 tweets in the last week, if a query was done to get those 50 tweets it wouldn't be guaranteed that the query would retrieve the 50 tweets, it could just retrieve 20 out of those 50.

Because of this limitations, when there was an attempt of getting data with this plan, few tweets were being retrieved because of this limitations and also because that there are users who make their account private so their tweets are not for public access and users might also not be active and just write for example five tweets in 7 days. So at the end of some queries few tweets were retrieved, which is an amount of tweets where conclusions can't be taken from.

At this point, a change of plan for the way data was being retrieved from Twitter API was needed.

It was then decided that data should be gathered based on a geolocation query, this way there would be a higher chance of having connections between the users that made the tweets since they were on the same area when they made the tweet. Based on this change there was also the need to change the way the edges of the network were being created between the users. After this change, an edge would be created on the network between two users if one user mentions another or one user retweets a tweet made by other user.

At the end, the way data was retrieved was with a geolocation parameter on the query to get the tweets. The geolocation that we inserted in the parameters of the query is in the center of London because it is a area with plenty movement of people and the main language is English which is the language that the sentiment analysis tool can analyze. This way of getting data had no problems in providing the amount of data needed to conduct a proper analysis.

Another detail about the Twitter API is that each query that is done can only get a maximum of 100 tweets so advantage was taken from this fact and each query was done by separating it from the previous query and next query by a certain time. This way users had time to keep writing to each other and more edges would show up on the network.

4 METHOD

The method executed to conduct the analysis of the data gathered and with the goal of proving the hypothesis had several steps:

- Getting the data from Twitter API and transforming it
- Calculating the sentiment score of each tweet present in the data
- Creation of the network
- Finding communities in the network
- Analysing the communities found and getting results from them

In the next paragraphs the steps will be explained. The sequence that the steps showed up previously is their respective order, this means that for example getting the data from Twitter API and transforming it is done before calculating the sentiment score of each tweet present in the data.

At the start, the data is extracted from Twitter API as explained on the chapter Data Extraction. From Twitter API, on each query a JSON object is returned with a list inside of it that contains the tweets from that corresponding query. Each tweet is also represented by a JSON object. So briefly, each query result from Twitter API was a JSON object with an attribute inside that was a list containing JSON objects inside which were the tweets. For simplicity reasons and as a first transformation, instead of saving the list of JSON objects (which represent tweets) we added one by one which made our file of data have one JSON object per line.

This file created will be used in the main program as it is right now, with one JSON object per line, but for the sentiment analysis tool we had to transform the data more so another file is created based on this data file.

The second file created is a file created based on the first file with the data collected from Twitter API. On this new file, the JSON objects from the first data file will be ready and the text, language of the tweet and user id will be taken from it and added in 3 different lines to the new file. This way at the end of the file there will be 'blocks' of 3 lines which each one of this 'blocks' represent a tweet. If a tweet is found to be a retweet then both the original tweet and the retweeted tweet are added to this file.

Since the purpose of this file is to be used as input on the sentiment analysis tool it is obvious why one of the lines is the text. About the other 2 lines, the language is included since vaderSentiment only supports English then the tweets which are not in English need to be excluded. The user id is included so the main program can keep track of who did that tweet.

After this is file is created, it is time to use it as input in the tool vaderSentiment which is described in the chapter background. The output of the tool is a new file which replaces the text of each tweet present in the input file with a small JSON object containing the classification of the text.

At this moment all the preparing steps before the execution of the main program are done and so it is time to execute it. The main program starts by creating a network using the package networkX (described in detail in the background chapter). The nodes of the network represent an user and the edges between users are done when a user mentions another user in a tweet or when a user retweets another user. The first data file is used for this process and at the end of the creation of the network we will have a network similar to this one which is the output of one of our tests:

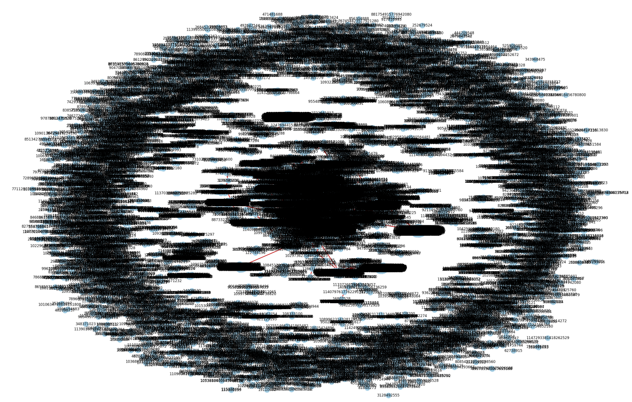

Figure 1: Network created by the main program.

Only the users with at least one edge to another user are added to the network since for our study propose those users are insignificant since they can't influence the sentiment of another user present in the network and by doing this the creation of the network is faster.

After the creation of the network as previously described, the file that contains the sentiment classification of each tweet is checked and the average sentiment value of each user present in the network is calculated. For example if a user wrote three tweets, one having 0.5 sentiment score, other having 0.1 sentiment score and the other one having 0.3 sentiment score then the sentiment score of the user would be (0.5+0.1+0.3)/3=0.3 .

At this point the network is created and each user has an assigned sentiment score so we can move to the next step which is finding the communities. Two different community finding algorithms were applied on the network because by using two algorithms the results at the end by using one algorithm can be compared to the results obtained at the end by using the other algorithm. With two different results at the end there is the possibility to confirm that the results from using a algorithm are confirmed or denied by another different algorithm one. If the results of one algorithm are confirmed by the other then we have a strong proof that the choice of the algorithm to find communities does not influence the final results and that our conclusions are not influenced by it.

The two community finding algorithms used are the Louvain method for community detection and the Girvan-Newman algorithm.

After applying this two algorithms, we now have for each of the algorithms the communities that were found. Each community contains a set of nodes and each node represents an user so it is known which users belong to which communities.

By knowing this and based on the sentiment score that was attributed before to each user, the number of communities that the majority of the users inside share the same sentiment can be calculated. This is firstly done by splitting the users in three categories based on the sentiment score they have. From -1 to -0.5 the users are considered to the sad, from -0.5 to 0.5 the users are considered to be neutral and from 0.5 to 1 the users are considered to be happy. Then every community is checked and the numbers of the communities where the majority of the users are sad, neutral or happy are calculated.

5 RESULTS

Three different datasets were used to have a good amount of outputs to be analyzed. One dataset had around 6500 tweets, other had around 10000 tweets and the last dataset that was used had around 15000 tweets. All the datasets took a different time to process, the dataset containing around 15000 tweets took more than 10 hours to run due to the fact that the Girvan-Newman algorithm is not efficient for big amounts of data. Here are the 3 image outputs to demonstrate how the communities that the algorithms found look like:

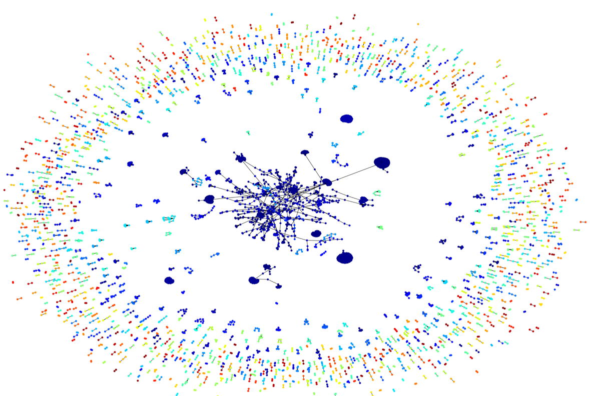

Figure 2: Around 6500 tweets dataset

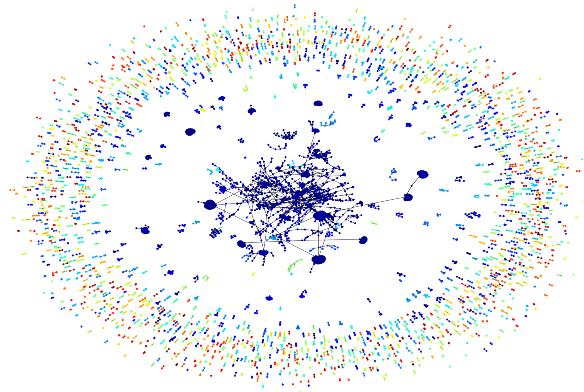

Figure 3: Around 10000 tweets dataset

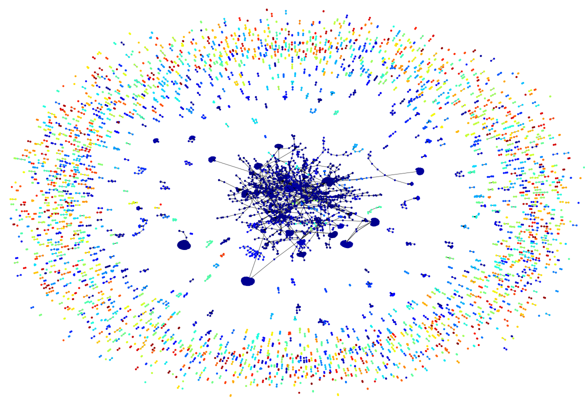

Figure 4: Around 15000 tweets dataset

As it is shown there are a lot of small communities but also some big ones. The colour of the node being the same to the colour of another node doesn't necessarily mean that the two nodes belong to the same community as one might think by looking at the images.

Another output that the program produces are the values in the tables presented below. In one of the tables for a community to be coherent at least 50% of the users inside the community need to share the same sentiment while on the other table at least 80% of the users inside the community need to share the same sentiment. Two different values (50% and 80%) were used in order that a comparison of results was possible. The results are really good at the 50% table but it is a low value to be strong enough to prove our hypothesis. On the 80% table the results are good and the value is high which makes it strong enough to have a strong proof of our hypothesis.

Number of tweets:	6500	10000	15000
Louvain method:			
# communities	1416	1937	2259
# positive communities	364	504	605
# negative communities	221	219	256
# neutral communities	564	777	903
# not coherent communities	267	437	495
Girvan-Newman algorithm:			
# communities	1453	1962	2276
# positive communities	368	511	610
# negative communities	238	221	257
# neutral communities	574	791	910
# not coherent communities	273	439	499

Coherent communities have at least 50% of the users sharing the same sentiment

Number of tweets:	6500	10000	15000
Louvain method:			
# Communities	1416	1937	2259
# positive communities	266	367	444
# negative communities	166	171	196
# neutral communities	402	539	609
# not coherent communities	582	860	1010
Girvan-Newman algorithm:			
# communities	1453	1962	2276
# positive communities	267	370	447
# negative communities	177	171	197
# neutral communities	404	543	611
# not coherent communities	605	878	1021

Coherent communities have at least 80% of the users sharing the same sentiment

As it is shown on the tables, the table with at least 50% of the users sharing the same sentiment has a clear majority of coherent communities and since 50% is a low number for our hypothesis the numbers of the table won't be analyzed in detail but are part of this report to have a comparison between the two tables.

On the second table, 80% is a high number and should be strong enough to get at least a good proof of the hypothesis. On the Louvain method and on the dataset with around 6500 tweets 59% of the communities shared the same sentiment, on the dataset with around 10000 tweets 56% of the communities shared the same sentiment and on the around 15000 tweets dataset 56% of the communities shared the same sentiment. Overall the average was that 57% of the communities had a majority of the users sharing the same sentiment.

On the Girvan-Newman algorithm and on the dataset with around 6500 tweets 58% of the communities shared the same sentiment, on the dataset with around 10000 tweets 55% of the communities shared the same sentiment and on the around 15000 tweets dataset 55% of the communities shared the same sentiment. Overall the average was that 56% of the communities had a majority of the users sharing the same sentiment.

From the results we can conclude that both the algorithms have a similar result which means that there is no influence of the community finding algorithm on the results and the results in all the datasets were close and even equal on the last two which might mean that the bigger the dataset the smaller the variation of this value. The overall values of the results show that a majority of the communities found have 80% of the users inside them sharing the same sentiment.

Since different datasets were analyzed and different community finding algorithms were used it can be concluded that the majority of the users inside the communities having the same sentiment can't be a coincidence and therefore our hypothesis that sentiments propagate over users gets a strong proof that it is true.

It is also worth mentioning that this is a subjective topic and with some inaccuracies. One reason for inaccuracies to exist is that is really hard for a computer to calculate sentiments present on a text, it can't detect irony or sarcasm for example. The tool used for the sentiment analysis has a dictionary to assign scores to specific words, as previous explained, but it might not contain all the words that express a sentiment and even if a word is there it might be wrongly written on the text analysed and the tool won't detect it as a word with sentiment.

To explain the subjectivity of the topic we mention that although users interacted directly with each other and therefore were able to influence others, nothing guarantees that their sentiment was influenced by that direct interaction.

Despite this facts, we are confident that this is a strong proof that supports our hypothesis.

6 CONCLUSION

After getting different datasets, building a network out of them based on direct interactions between the users, finding communities in the network and then analyzing them, it is our belief that we found a strong proof that sentiments propagate over twitter from users to another users since in all the different outputs there was a majority of the communities having the majority of the users sharing the same sentiment which is really improbable to be a coincidence.

REFERENCES

[1] Daniel A. Schult Aric A. Hagberg and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)* (Aug. 2008), 11–15. http://conference.scipy.org/proceedings/scipy2008/SciPy2008_

proceedings.pdf

[2] Belem Barbosa. 2017. *Happiness in marketing*. 75–90.

[3] C.J. Hutto and E.E Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text? *Eighth International Conference on Weblogs and Social Media (ICWSM-14)* (June 2014). http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

[4] Girvan M. and Newman M. E. J. 2002. Community structure in social and biological networks. *Natl. Acad. Sci. USA 99* (2002), 7821–7826. https://www.pnas.org/content/99/12/7821

[5] Renaud Lambiotte Etienne Lefebvre Vincent D Blondel, Jean-Loup Guillaume. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008* (2008). https://arxiv.org/pdf/0803.0476.pdf