Social Media Influence Analyzer

Mohammed Z. Guniem  
Department of Computer and Electrical Technology  
*University of Stavanger*Stavanger, Norway  
mghunime@yahoo.no

In a time where online social media has a significant influence on both individuals and groups of people in our modern societies, there is a need to have a digital tool that helps gather and construct data from online social media, then use this data to detect influence between individual users by detecting, measuring, and classifying the strength and topic field of influence in the process. This project proposes a method for influence detection, then build a technical solution on top of it to enable analyzers of online social media to have a visual and scientific understanding of the influence flow in social media network.

Keywords: Social-media, influence-detection, influence-topic-classification, text-classification, graph-centrality, graph-node-ranking, IT-system-architecture, logging, caching, HTTPS-Basic-Authentication.

Motivation

Upon the rise of the digital revolution through the last two decades, people around the world are no longer limited to the constraints of place and time to socialize with each other. The newly introduced concept of digital media has transformed the way our society function and socializing no longer requires the physical presence of society members. As a result of the introduction of multiple social media platforms, people from all over the world can now engage in local, national, and global events, participating in society and expressing themself in an open arena where physical boundaries do not stand in the way.

Today and after a very short time of experiencing the advantages of social media platforms, our society has become almost totally dependent on such platforms, and most social events and happenings does not pass away from being recorded and discussed in the wide arena of social media. This effect generates a huge amount of valuable data that has a big potential of revealing the type and strength of social influence between society members and opens for many useful applications in multiple fields.

The most obvious application from social data is understanding how social media is used as a tool to mobilize groups of people in controversial social events such as political elections. The serious allegation of Russian interference in the US presidential election in 2016 is one application of analyzing social influence on social media.

Furthermore, by mapping and visualizing social influence between users on social media, we can speed up and improve the detection of fake news and other illegal activities on social media, and by removing their damaging effects on multiple social environments, we can create a healthier society that benefits all its members.

Social influence is also highly valuable for commercial use, as many companies are interested in detecting different types of social influence to reveal new marketing trends and allow businesses to develop more specialized marketing strategies and customized products which often increase the competition in economy and generate more values for companies and their surrounding societies.

These were some applications that can benefit from analyzing influence between users on social media, and there is still both uncovered and undiscovered areas where understanding social influence is highly crucial for the purpose of the application.

# **Introduction**

Data from social media has a great potential in revealing how strong the influence is between different users, just like in real life every action a user commit and how society members react to this action can serve as an object for analysis which helps in drawing a big but rather detailed picture of how users influence each other across many societies and fields.

The aim of this research is to establish a ground foundation for extracting information about user activities on social media and use such information to detect social influence between network users. Such foundation is desired to make up the core of a future technical solution that enables social media analyzers with little or no technical experience in data processing and visualization to perform social media analysis on regular periods with a continuous timeline.

To serve this purpose, we start by determining the common characteristics in available user functionalities on the most popular social media platforms, then produce a model for data structure based on these similarities in user functionality, and by taking a starting point of common user functionalities, we increase the flexibility of this research to be applied to as many social media platforms as possible, and perhaps combine results from several platforms in one single analysis if needed.

After establishing an agreement on the data model to be used for collecting and storing crawled data from social media, we dive into the main core functionality of detecting social influence between network users. Multiple techniques of detecting and scoring social influence will be implemented to fit the different needs of a final analysis. The desired result is a user-oriented influence graph where each node represents a participating user, while each edge between two given users represents the influence between them with respect to direction of influence, and holding the score of influence strength along with its classified field of influence whether it is in sports, politics, or economy etc.

Following the previous effort, we evaluate the performance of the influence graph model and go through test results from both dummy and real-life data using crawled data from a rising social media platform called “Reddit”. We will then try to highlight the most interesting and useful features of the produced influence graph and push its power of detecting influencers and their area of influence to the limit. The final two processes of test and evaluation are together a vital step to confidently rely on the quality of the produced model of the influence graph by ensuring its informative capabilities in social analysis.

Furthermore, a technical solution is to be designed and implemented to work hand in hand with the theoretical approach and function as a possible practical implementation used as a proof of concept and as a helping tool in testing and evaluation. During the process of designing and building this technical solution, many important aspects of data protection, reliability and availability are discussed and dealt with to improve the ecosystem of this application.

This was a brief introduction of the upcoming research in a nutshell, but first let’s take notes and learn from some interesting pre-attempts in studying influence between users of online social media.

# **Related Works**

Among the community of data science, a wide variety of studies has focused on extracting information from online social media, and a great amount of effort has been dedicated to studying social influence between users to better understand the behaviour of individuals for many purposes. Research of social influence takes different forms and vary in size and scope, while some researchers take on the very fundamentals of detecting social influence, others dive through it to reveal details such as a specific influence or hidden behaviour patterns on different levels. In this section, we explore some related work in the field of social activity on online social media and try to get an inspiration that helps direct the effort of this research in the right path.

## **Measuring Influence Between Users of Online Social Media.**

A good fundamental approach is described by a social network analysis carried out by Y. Guo, J. Cao & W. Lin. The fellow researchers divide the influence evaluation models into 2 main categories; the first category is based on network topology which measures social influence between different users by considering the user degree, shortest path, and some random walk characteristics, while the second category bases the influence between users on their interactions through different activities organized in a tree-like structures containing submissions and multilevel comments. However, and despite the reasonably good classification and overview, the published paper of this research lacks some proven results of an experimental approach [1].

## **Data-driven Influence Learning.**

A short but rather interesting experimental and mathematical approach is introduced by a paper on Data-driven Influence Learning in Social Networks published by F. Wang, W. Jiang, G. Wang & D. Xie. In this paper, the process of influence diffusion is divided into two parts: the launcher (influence strength) and the receiver (influence threshold) which can generate an accurate and finer grained influence diffusion model according to this research. [2]

Furthermore, the researchers highlight the importance of having a solid criterium when scoring the strength and threshold properties of social influences. Another important acknowledgment is the difficulty and complexity associated with detecting influence relationships between users as a by-product of big datasets that usually include a considerable amount of noisy or less important datapoints, making it essential for any algorithm used in learning and testing the influence models to perform a minimal scan over the data in the most efficient way possible.

## **Alternatives of Information Gathering.**

Most well-known providers of social media platforms assist developers and data scientists with instructions on how to crawl their platforms by offering multiple endpoints and methods that can be used for gathering data for analysis.

Multiple researchers spot the light on this initial aspect of gathering data from social media platforms. A significant research is one that mainly describes the alternative of Pushshift Reddit Dataset by J. Baumgartner, S. Zannettou, B. Keegan, M. Squire and J. Blackburn. [3] This research paper offers an undirected but also claimed to be a more efficient and flexible way to gather data from the “Reddit” social media platform, in comparison to using the official Reddit API endpoint.

It also gives an excellent brief description of the FAIR data[[1]](#footnote-1) principles which is highly relevant when choosing the source of data especially when it comes to accessibility and findability.

Another advantage of this research is its extension in discussing a series of the other major alternatives for gathering data from “Reddit”, highlighting their strengths and weaknesses in a constructive manner.

## **Topic Detection in Socical media platforms.**

As mentioned in the introduction, we are set to determine the category of a detected influence between users, this opens for the use of artificial intelligence for the purpose of classification between different topics where a certain user activity might fit in. Inspiration on possible solutions for this task can be obtained from a research about annotating and detecting topics in social media forum and modelling the annotation to derive directions carried out by B. Athira, J. Jones, S. M. Idicula, A. Kulanthaivel and E. Zhang. [4]

A practical case study from an online health community was represented to give a good introduction of data pre-processing and cleaning, then preceding to construct a reasonable mathematical approach in the training and testing of a machine learning model to be used for the purpose of topic classification.

Another contribution of this research is the use of various deep learning algorithms to classify posted content such as CNN[[2]](#footnote-2), LSTM[[3]](#footnote-3) and BiLSTM[[4]](#footnote-4), all in which enable the researchers to achieve a promising F1-score[[5]](#footnote-5) of about 0.75 to 0.80 in topic classification accuracy.

Furthermore, the above research offers a solution for a much-needed ability to minimize the amount of training data and dealing with the negative effects of label imbalance in a training dataset, then constructing a convincing conclusion after carrying out a process of well-performed testing and evaluation, where metrics of evaluation are carefully examined and explained in a good scientific approach [4].

## **Study Case Alternatives of Online Social Media.**

Determining which social media platform to crawl under testing and evaluation of a new modelling approach is important to produce a flexible influence model that can be used for analysis of as many social media platforms as possible, this is why it is desirable to work with real-life datasets gathered from a digital media platform that shares common user functionalities with as many popular social media platforms as possible, examples of such functionalities are posts or submissions, comments, and upvotes or commonly known as likes.

A social media platform that satisfies all these user functionalities is “Reddit” which is examined by the research with the title “Information and Social Analysis” carried out by T. Steinbauer at the University of California, Santa Barbara. [5]

Steinbauer starts off with a brief but very constructive comparison between the most popular sites for social news with Reddit included. The core of Steinbauer’s research lays in his analysis of subreddits, submissions and comments on the virtual platform of Reddit, this analysis helps explaining why Reddit should be used in evaluating the performance of an influence model and its ability to view the most influencing users in a social media platform. The reason for this is Steinbauer’s detailed analysis on which subreddits seems to have the most of user’s activity, and in addition his further construction of an example user-oritened influence graph that helps showing which user has the highest influence based on the user’s interactions through comments.

However, submission authors are not included in the dataset of the constructed influence graph, making this influence model less reliable if ignoring the often-significant role of posters in generating discussions on social media. Another downside of Steinbauer’s modelling of an influence graph is the limitation of not using any other criteria than user interaction through comments, such as the upvote score or number of thread or descendant comments posted on other comments or submissions.

Although Steinbauer has introduced a detailed result overview of his evaluations and analysis, there is still a question mark on the technical details because algorithms that has been used for producing the model of the influence graph are not provided to the reader in satisfying details.

# **Reddit as a case study social media**

There exist a wide variety of popular social media platforms and most of them are constantly gaining popularity among users from all over the world.

The following figure shows the market share of the top 7 most popular social media platforms during the last decade from 2010 to 2019, where market share data is obtained from statcounter.com which claims to base its statistics about digital market shares on a sample exceeding 10 billion pageviews per month. [6]



Figure 1, Worldwide market share of top popular social media platforms from the beginning of 2010 to the end of 2019 [7]

Although Facebook is the definite leading social media platform, it is still possible to observe a competition in popularity when looking at the next 6 platforms below Facebook, with Reddit having a popularity corresponding to all other social media platforms that are less popular than Reddit.

A normal side effect of a more popular social media is the large amount of data users generate on such platforms, which slows the process of extracting data from such platforms, and although data from a more popular media often has a higher integrity, it is important to keep a balance between data integrity and easiness in findability and accessibility. In this research, we try to compensate between these two factors by choosing a medium-sized social media platform for use during testing and evaluation of the influence graph model.

The market share of Reddit does nearly equal as the sum of at least 14 other social media platforms and all of these are reported to be below reddit in popularity including well established platforms such as LinkedIn and Instagram. This makes Reddit a suitable candidate to be used as a case study social media in this research, as Reddit offers our desired moderate balance between network size and easiness to crawl.

In addition, many of the most popular social media platforms tends to specialize in a certain area or field of social activities such as LinkedIn for professional life, and Facebook on the other hand mostly used for private and personal socializing, some digital platforms combine aspects from both areas such as the so-called digital news platforms that offers its users an opportunity to interact with each other in many aspects of socializing like professional and personal life combined. Reddit is considered as one those digital news platforms which is still gaining popularity and increasing in content since its launch in 2005.[[6]](#footnote-6)

A user on Reddit can create or join a group, make a submission on any group and comment on any submission or comment made by other users. A user can join a group, but it is not obligatory to join a group to be active in them or read their content. These groups are called subreddits and tend to specialize in a certain topic of interest in society, and for many users it is seen in a way that is somehow like reading the newspapers which is often divided into pages for multiple areas of concern such as politics, economy, or sports. The high separation between topics of interest in Reddit makes this platform ideal for testing how well an influence detecting algorithm can discover and classify different types of influence between users.

Reddit differs from other social media platforms in the sense that Reddit attracts users by their interest in topics and events in their social surroundings, while other social media often relays on the social affiliation of a future user. However, and on the other hand, many other social media platforms share a lot of common user functionalities with Reddit, such as groups, submissions, and comments.

This high similarity between Reddit and most popular social media platforms, along with Reddit’s ability to separate users into multiple different social groups, makes Reddit very suitable as an evaluation study case for this research as common functionality increases the modelling flexibility to be used on other social media platforms in future analysis, and its separation of social environments in groups serves the purpose of comparing the predicted type of social influence between users to the actual definition of the group where the interaction between users has occurred to give us an idea of how well our influence model is classifying topics of social influence.

Although Reddit is a user-oriented platform, its users often prefer to be anonymous, which is useful when presenting results without having to worry about neutrality issues, but Reddit’s users can also choose a username that can be used to identify them personally if they desire.

Another good reason for choosing Reddit as a study case is the highly developed endpoint crawling API which is very object-oriented and offers a wrapper library for the Python language. This eliminates the bother of dealing with HTTP requests and latency issues, as all of this is taken care off in the background of the Python Reddit API Wrapper. The Wrapper is free to use but it requires a registration which once done offers no restrictions on how often Reddit is crawled, unlike crawling by adding “.json” to the URL which have its downsides such as limitation for under 100 submissions at a time, and the blockage of multiple requests from the same IP address as a prevention measure from Reddit to stop denial of service attacks. All these downsides are escaped by using the Python Reddit API Wrapper which increases the reliability and stability of data streams from reddit. In other word the PRAW[[7]](#footnote-7) python module satisfies the following FAIR data principles:

* Findability:

Once using PRAW, it is easy to find and retrieve data from Reddit no matter how detailed the data is.

* Accessibility:

As mentioned earlier a programmer does not have to deal with HTTP requests and latency issues as when using a traditional API endpoint, this makes the programming experience much easier allowing programmers to focus on the objective of their work.

* Interoperability:

A good documentation and maintaining history of the PRAW module [8], along with its popularity between programmers which gives it an excellent record of ability to integrate with different products and systems that uses it.

* Reusability:

PRAW is highly object-oriented in both query language and retrieval results. This is helpful for the usability for integration in different projects and technical solutions both in present and future technologies.

Based on the above four FAIR data principles and the previous analysis of user habits and possibilities on Reddit and other social media platforms, Reddit makes a good case study in the testing and evaluation process when we are seeking to detect user influence and their area of influence. We shall than design our ground data structure to adapt for the common functionality between Reddit and the most popular social media platforms as we will go through in the upcoming section on the ground data structure.

# **Definiing a ground data structure**

Flexibility of design is an important requirement of this research, as we aim for a future application of the influence modelling developed in this research on as many other social media platforms as possible, and although this might be difficult to achieve as a result of the wide variety of available social media and the different user-functionalities in them, we can still notice some common user functionalities between the most popular social media platforms such as LinkedIn, Facebook and Reddit, this common functionality is no accident, as these social media platforms most likely inspired from real life social interactions to begin with, which in turn is a natural advantage for our application.

After studying the available user functionalities in Reddit compared with these same functionalities on the most popular social media platforms, it is easy to see a big potential for developing a generalized data model that can be used to structure data crawled from any of the applying social media platforms. It is therefore important to consider the desired results of this research before establishing a ground model for data structure.

Social influence can be defined to be the ability of one society member to change the thoughts or behavior of another society member, and although this definition is simple, the complexity is hidden in the way social influence plays out in real life society. Some people get influenced without any big significant reaction that can be recorded and studied, such influence is said to be of a passive type of influence, an example of passive influence is reading a newspaper where the reader gets influenced without adding any additional comments to the content.

The main goal of this research is to use recorded data from social media to visualize the influence flow between a group of people in a social interaction. For this reason, we are going to look at active social influence where we would expect the person who get influenced to react by submitting an activity on the content of influence. This requires an activity-based model, where activities such as submissions and comments are considered as indicators for social influence.

The second requirement of this research is the importance of visualizing the flow and direction of influence between members in a social media interaction. For this purpose, we will be building an interaction-based model that is able to retain the origin and target of each detected and measured influence, which benefits the storage of influence direction and in the big picture can be used to visualize the entire flow of social influence between society members.

The model can initially be based on four different entities a user can create and interact to; these entities are:

1. Network

Which holds information about the crawled social media platform or a smaller segment of it, having this entity, makes it possible to study multiple social media platforms at the same time which increases the flexibility of design.

1. Group

A group contains a bundle of submissions posted by users of the group. It also contains information about a certain group in form of identification and other attributes such as the group ID and name.

1. Submission

A submission is posted by one user and is assumed to be in text format with the possibility of further extension to multiple other formats like images and links in future improvements. The submission entity has multiple useful information stored in its attributes such as the current number of- comments, and -upvotes on this submission, along with information about the author of this submission and other identity attributes.

1. Comment

The comment entity is very similar to the submission entity containing a body text, identification of author and information about the location of a comment in the comment thread. And therefore, it has an additional feature which its ability to be a parent and/or a child of other comments. This means that comments can be modelled as a tree data structure that can grow unlimited.

The figure below shows a diagram of an entity-relationship model that will make the base of our data structure further on in this research. In addition to the four entities explained earlier in this section, four relationships bind these entities together defining their relations to each other. A network can contain multiple groups, and a group can contain multiple submissions or posts, where users can either comment on those submissions or on other comments that is a child descendant of the comment tree of a certain submission.



Figure 2, The Entity-Relationship model which represents the ground datastructure of this research.

Most of the popular social media platforms contains the four identified entities in this ER-model[[8]](#footnote-8), although they might have a different name, form, or purpose such as a company page on LinkedIn or a user profile on Facebook, but both can be treated as groups just like Reddit groups as well.

The attributes of entities open for more flexibility as we might have the need to extend or shrink our ER-model in the future i.e., by not including an upvote attribute or by adding a reaction attribute to submissions and comments. But also, attributes are generalized to match the very common details about of these entities in between the most popular social media platforms.

Now that we have a ground Entity-Relationship model to base our data structure on, we can proceed into discovering influences between users based of the interactions between them using these entities and relationship roles established in this section. In the next steps, we will elaborate the different stages of our influence modelling process for detecting and extracting social influence from online social media, the algorithms of the upcoming influence modelling are expected to have multiple dimensions for revealing the strength and types of social influence between different users online.

##### References

[1] Y. Guo, J. Cao and W. Lin, "Social Network Influence Analysis," 2019 6th International Conference on Dependable Systems and Their Applications (DSA), 2020, pp. 517-518, doi: 10.1109/DSA.2019.00093. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] F. Wang, W. Jiang, G. Wang and D. Xie, "Data-Driven Influence Learning in Social Networks," 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), 2017, pp. 1179-1185, doi: 10.1109/ISPA/IUCC.2017.00177.

[3] Baumgartner, Jason & Zannettou, Savvas & Keegan, Brian & Squire, Megan & Blackburn, Jeremy. (2020). The Pushshift Reddit Dataset.

[4] Balakrishnan, Athira & Jones, Josette & Idicula, Sumam & Kulanthaivel, Anand & Zhang, Enming. (2021). Annotating and detecting topics in social media forum and modelling the annotation to derive directions-a case study. Journal of Big Data. 8. 10.1186/s40537-021-00429-7.

[5] T. Steinbaur, ‘‘Information and social analysis of Reddit,’’ inProc.TROYSTEINBAUER CS. UCSB. EDU, 2012, pp. 1–12. [Online]. Available:http://snap.stanford.edu/class/cs224w-2011/proj/tbower\_Finalwriteup\_v1.pdf

[6] "About Statcounter GlobalStats." https://gs.statcounter.com/about (accessed July. 12, 2021).

[7] Statcounter GlobalStats, Social Media Stats Worldwide, Jan 2010 - Dec 2019. [Online]. accessed from <https://gs.statcounter.com/social-media-stats#monthly-201001-201912-bar>

[8] "PRAW: The Python Reddit API Wrapper." https://praw.readthedocs.io/en/stable/ (accessed July. 12, 2021).

1. FAIR data are data that meet principles of findability, accessibility, interoperability, and reusability, <https://en.wikipedia.org/wiki/FAIR_data> [↑](#footnote-ref-1)
2. Convolutional neural network, <https://en.wikipedia.org/wiki/Convolutional_neural_network> [↑](#footnote-ref-2)
3. Long short-term memory, <https://en.wikipedia.org/wiki/Long_short-term_memory> [↑](#footnote-ref-3)
4. Bidirectional LSTM, <https://paperswithcode.com/method/bilstm> [↑](#footnote-ref-4)
5. A measure of model’s accuracy on a dataset, <https://deepai.org/machine-learning-glossary-and-terms/f-score> [↑](#footnote-ref-5)
6. Reddit is a social news aggregation, web content rating, and discussion website, <https://en.wikipedia.org/wiki/Reddit> [↑](#footnote-ref-6)
7. Python Reddit API Wrapper [↑](#footnote-ref-7)
8. Entity-Relationship model [↑](#footnote-ref-8)