Social Media Mining and Language

Zain Rajwany
Student / Indiana University
6842 Fieldstone Drive
Burr Ridge, IL 60527
zrajwany@iu.edu

Abstract

This paper will cover the basics of social media mining, and in particular describe how it pertains to linguistics. Social media has grown so much in our society, people are sharing and making content at a huge rate. Organizations can collect this data and analyze it to come to conclusions about the behaviors of the users of social media sites. The results of social media mining can be used for marketing campaigns targeted at certain users. A large part of this paper will go over the basis of social media mining, and how it uses the behavior of society to capture trends across the world.

1 What is Social Media Mining?

The reach of the internet can be hard to comprehend at times. Millions and millions of people spend countless hours online, and many of those hours consist of sharing, communicating, and creating data. Social Media has become a source of ”big data,” which is basically a field that deals with data sets that are very complex. This data is different than what you may typically expect, in that there really isn’t a structure to the relations between users of social media.

According to Zafarani (2014) social media mining is ”the process of representing, analyzing, and extracting actionable patterns from social media data.” Zafarani (2014) Organizations collect data and info from social media users, and they analyze it to come to conclusions about the behavior of said users. Social media provides us with the unique way to study how people interact online and how communities form within these social media sites.

The idea of mining into this huge amount of data requires algorithms that are able to decipher it. When companies can collect data using these algorithms, they can analyze it to possible target specific groups of users in marketing campaigns, or to observe the attitudes the public has toward certain people/things. We can also observe interactions between people as well. As Zafarani et al. states, ”Social theories and social norms govern the interactions between individuals and entities. For effective social media mining, we collect information about individuals and entities, measure their interactions, and discover patterns to understand human behavior.” Zafarani (2014)

2 Challenges in Social Media Mining

”Big data” is, well, big. If an organization wanted to make suggestions based on a user’s history depending on the things they may have typed, it would be difficult to center in one what little data would be available for that specific person in a specific context. To add on to that point, ”how can we ensure that our findings obtained from social media mining are any indication of true patterns that can benefit our research or business development?” Zafarani (2014) It’s important to know what the true patterns are when it comes to what the customers are saying on social media, so your business can grow.

Data mining includes removing ”noisy” or irrelevant data. However, lazily overlooking this data can cause you to miss potentially important information.

3 Preparing Data

Processing large amounts of data is almost impossible, which brings the importance of sampling. A sample is a small random part of the data that is analyzed instead of the entire thing. Sampling techniques must be used correctly so that the results we get are similar to what we would find from the whole data.

4 Unsupervised Learning

Unsupervised Learning is a category of data mining algorithms. In unsupervised learning, the idea is to find similar instances in a dataset, and group them together. By grouping them together, patterns emerge. The example given by Zafarani et al. shows that unsupervised learning is used to find and identify events on Twitter. ”Tweets can be grouped based on the times at which they appear and hence, identify the tweets corresponding real world events.” Zafarani (2014)

By grouping the language that is similar across tweets as well as the time, we can relate them to what is happening in the real world.

5 Communities

There are numerous reasons to study communities in social media mining. The main reason is that individuals who share interests and beliefs tend to form groups, especially on social media sites like Twitter or Instagram. Whereas an individual user has very unique and therefore not important data on their own, groups can show us a potentially ”global view of user interactions”. Zafarani (2014) If there is a certain group of people who hold a specific view towards a person or company, the similar language used by this group can be seen. Some behaviors can only be observed in a group, as an individual can be persuaded to think or act differently while a group tends to hold strong opinions together. As we’ve seen, social media is a hotbed of groups who hold different opinions on the same subject, with debates happening all the time.

6 Information Diffusion

Information diffusion is defined by Zafarani et al. as ”the process by which a piece of information (knowledge) is spread and reaches individuals through actions.” Zafarani (2014) An example they give for this was a viral tweet by the Oreo company. During the infamous power outage in the Super Bowl a few years ago, Oreo gained attention that would normally cost millions in an ad in a simple tweet with also drew attention to their product. The tweet from Oreo ”diffused” into a huge number of people.

Diffusion consists of three elements: the Sender, the receiver, and the medium (the communication process between the sender and receiver).

Zafarani et al. lists 4 general types of information diffusion in their book: herd behavior, information cascades, diffusion of innovation, and epidemics.

Herd behavior is when someone looks at the behavior of many other people and decides to act in a similar way. It’s probably no surprise to find out that just like in the real world, people can make decisions by conforming to peer pressure and/or social norms. Herd behavior is defined by Zafarani as individuals consciously making decisions aligned with others. Zafarani (2014) We can see this when an event is trending on Twitter, people will look at the language being used by others with high like and retweet numbers, and consciously choose to follow the same opinion.

Information cascades are just the act of individuals observing neighbors and making decisions based on them. An information cascade happens as info gets passed through familiar friends. This is most common on sites like Facebook where you can share posts made by a friend of yours to your friends. Since information cascades happen only through a network of immediate friends, less information is available as opposed to herd behavior.

Diffusion of innovations shows us how something viral (music video, product, meme) spreads through the population of a social media site. People’s words about these certain viral things are what make it easier to track. The theory of diffusion of innovations tries to answer how innovations spread, as well as showing the reasons why it happens, and how fast ideas can spread over social media through the things people say and share.

Epidemic models are similar to diffusion of innovation except that people do not decided whether of not they become ”infected.” The name stems from the word epidemic being used in its usual context - in which a disease spreads throughout a population of being that hold said disease. When looking at global patterns, epidemic models work better as they allow us to track social media trends as people learn about global events through the internet.

7 Influence and Social Similarity

Social media allows people to connect from anywhere as opposed to only meeting people who live around you. When people connect, there are patterns that arise. One of these patterns is called ”social similarity,” or ”assortativity,”. Social Similarity networks are defined by ”nodes” that are connected to other similar nodes more frequently than ones that are different. To relate this to social networks and language, ”This similarity is exhibited by similar behavior, similar interests, similar activities, and shared attributes such as language, among others.” Zafarani (2014) Companies can study these networks on social media to gain insight into interactions between large groups of friends. This is also known as homophily, which is the pattern of similar individuals becoming friends.

The force of influence is one of the causes for these networks. Unlike homophily, influence is one person/figure affecting others, instead of two comparable people connecting with each other. An individual can affect others in a way so that the now influenced group of people act more similarly to the original figure. The PageRank discussed in class in an example of one way to measure influence by prediction. As Zafarani et al. describes it, ”we can assume that the gregariousness (e.g., number of friends) of an individual is correlated with how influential she will be.” Zafarani (2014)

Another way to measure influence is by observation. Celebrities have influence over a large group of people, and the manner in which they speak/what they say resonates with their followers. Something else that companies are especially interested in are looking at how people talk about their products. When people buy products and post about it, it can make that product more valuable. Companies can take advantage of this by paying people with large influences to show off a product. This gets people mentioning their product through social media, as the information gets diffused through the online population.

8 Recommendation

Every day people around the world make decisions online. They can vary from what goods and services to purchase to adding new friends on Facebook. These decisions, especially when deciding on something like buying a common item, have many options. This combined with the limited knowledge of the buyer can be frustrating, as one wants to get the best item they can. A search engine can help find what a person is looking for, but it’s difficult to find results that match exact tastes, and can be dependent on what exactly you type into that search bar.

Algorithms called recommendation systems are developed to solve this problem. Recommendation systems ”are developed to help individuals decide easily, rapidly, and more accurately. These algorithms are tailored to individuals tastes such that customized recommendations are available for them.” Zafarani (2014) A search query looking for good books to read will give the same results for a 10 and 50 year old who may have different tastes. Recommendation systems usually look through your history, past purchases, your social media profiles, and information from your friends and peers to customize your search.

Recommendations can be used in a social media context to show users potential people they might know to send a friend request. On twitter, you can see recommendations of who to follow based on other people they follow, as well as possible similar things you tweet about.

There are some obstacles that occur when building recommendation systems. If the user has no historical data, how can an algorithm guess what that user will like? Sites can ask users to answer questions pertaining to a topic or require users to put interests in their profile to combat the problem. Another problem that is more difficult to solve is the privacy concern. The more info a system gets from a user’s profile, the better the system can work. However, users are understandably concerned with revealing information about themselves, and so this continues to be a challenge in this area.

9 Brand-Related User-Generated Content on Twitter

In a study done by Xia Liu et al., researcbers ”utilized social media mining techniques to gauge users perception of a variety of common brand names.” McCourt (2018) Text was mined and analyzed on 1.7 million different tweets for 20 brands in 5 industries. The industries were fast food, department stores, footwear, electronics, and telecommunications.

Social media has become a vital part of marketing communication. Users on sites like Twitter and Facebook can interact with brands by sharing positive or negative experiences. There’s a lot of potential value in this data with regards to advertising and communicating with customers.

Twitter has hundreds of millions of active users, which would make it impossible for a person to realistically go through the relevant ones. Using Twitter's Streaming Application Programming Interface (API) and algorithms written in Java, Liu et al. were able to collect around 10 million tweets. Twitter's API ”allows users to pull tweets off of Twitter according to certain keywords.” McCourt (2018) After processing this data, about 1.7 million tweets were kept.

The product, service, and promotion appeared to be the three topics that users on Twitter were the most interested in. When talking about the product, consumers tended to focus on the quality. An example the journal gives is a tweet collected that praises Burger King for their fries. When it comes to promotions, ”Users also tend to compare quality and innovation of products among competitor brands in their tweets...Furthermore, consumers are interested in current news or trends related to the brands.” Liu (2017) Competition between quality of fast-food restaurants was prevalent, and celebrity news was found to be influential, particularly in footwear.

The study also covered the attitude of tweets in their data, classifying them as positive, negative, or neutral. It was found that the telecommunications industry had the highest average percent of negative tweets, and in particular Comcast was the worst offender, with 66.7 percent of tweets about the company being negative. In addition, the percent of negative tweets in general was found to be much higher than positive ones. Negative tweets are bad for a company's image, and information like this helps companies see how they are perceived by customers.

The next issue was how the researchers could identify specifics. What products and services were the customers complaining about? Comcast was chosen for this as they had a large percent of negative tweets. Across a year, the phrase ”worst customer service shows up in every month, and it tends to account for between 20 percent and 30 percent of all tweets.”” Liu (2017) Other complaints dealt with billing issues and internet outages. This is useful because a company can use this analysis to implement strategies to solve the biggest problems facing that company.

The same analysis can be done for positive tweets as well. The data shows that the internet service of Comcast was the biggest positive aspect. However, the percentage of positive tweets were significantly smaller.

The findings for this project ”provide brand managers with actionable insights in targeted advertising, social customer relationship management (CRM), and brand management.” Liu (2017)

10 Advertising Based on Geotagged Social Media Data

This next research project also used twitter, but this time the goal was to optimize ”digital out-of-home” (DOOH) advertisements in the London Underground. An example of a DOOH ad is ”a digital billboard programed to change the advertisement on display after a specific period of time.” McCourt (2018) The research done shows how interests from locals can be mined by using tweets that were geotagged, and this information can be used for advertising.

Because of the ability to see what time and place a person tweets from, Twitter is a great source to understand interests of the residents in an area. In this project, the researchers were interested in the Tweets around each station of the London Underground, specifically focusing what topics at what times. ”The information can be used to determine the dominant interests for a particular location and time, or alternatively, identify the best locations and times for particular target audience instead.” Lai (2017) The London Underground was chosen because of the large number of people traveling through, and advertising is provided through digital screens.

After collecting data for over a year, and doing the necessary extracting, analyzing, and compiling, it was shown what people were tweeting about on specific times of the day on both weekdays and weekends. As an example, about 40 percent of tweets from 6pm to midnight on weekends at a certain station related to music, while over 40 percent of tweets from another station related to music on weekdays at the same time. Using this information, the researchers conclude that ”Geotagged Tweets could, therefore, be a useful tool for estimating variations in public interests across space and time,” Lai (2017) although it was noted that Twitter users are not representative of the entire population.

References

Juntao Lai. 2017. Improved targeted outdoor advertising based on geotagged social media data.

Xia Liu. 2017. An investigation of brand-related user-generated content on twitter.

Abby McCourt. 2018. Social media mining: The effects of big data in the age of social media.

Reza Zafarani. 2014. *Social Media Mining: An Intoduction*. Cambridge University Press.