Daniel Crawford

CS6795 Cognitive Science

24 April 2023

Milestone 4

I. Abstract

Social media has dominated digital communication since its arrival. There are unprecedented levels of communication that take place which not only change the way humans communicate, but the processes that affect this way. Semantic drift is a highly influential way of understanding the changes of meanings of words. With the social media storm, there is no doubt an effect on the behaviors of semantic drift. This project seeks to develop an understanding of the cognitive science underpinning to the manner in which social medial effects semantic drift. Conceptual framework is presented along with competing computational models. The models, as well as empirical analysis were conducted on a dataset featuring a month of twitter data regarding Covid-19. There Appears to be a highly connectionist view of social media and its effects on semantic drift.

II. Introduction

The grand rise of social media has shifted human communication in a more impactful way than perhaps anything since the dawn of writing. Indeed, we see the escalation form conversations with those exclusively in one's immediate surrounding to the entire digital world in mere decades. From this, we can only expect a proportionally significant change in the way language changes accordingly.

Language changing is a natural phenomenon that has been in place for thousands of years. But, social media has not. There is still much research being done in this field of course, but it appears to be the case that social media is changing the way languages change. If we consider the way that constant digital communication alters the meaning of words due to the fact that one is not crystallized in their linguistic background, it is most reasonable to expect that the meaning of individual words changes likewise.

If it is the case that social media is changing the way that languages change, through the altering of semantics, there should be a cognitive science framework to capture this. Cognitive science, which is the field that investigates at its deepest level the way human cognition processes occur, has designated frameworks for understanding human language change. With these pillars of thought, this project will seek to use social media data to computationally investigate the manner in which social media affects semantic shift and importantly, how this can inform theories of cognition.

A. Linguistic Change

Similar to the way life changes over time and through generations, the languages of humans do the same. In fact, languages are always mutating. The speakers of the language who seek to represent novel ideas and expressions make use of the communication they possess, and develop new ways of thinking, feeling, and communicating. While there is a vast array of manners in which languages can accomplish this change, the goal will be to understand the way that meanings of lexical items change.

B. Lexical-Meaning Pairs

First, it is important to establish the framework in which the investigation will be conducted. Define a lexical meaning pair as a lexical unit that is associated with any meaning.

This relationship is not necessarily one-to-one. There could be a many-to-one relationship, such as in the case of friends, buddies, chums, which all refer to peers with amicable standing, or a one to many, such as in the case of set, which is famous for its large number of definitions in American English.

It is important to note that we are not exclusively referring to words here. A lexical unit could be a collection of words, such as a phrase or idiom that has a socially concrete meaning, such as the description of a person as a 'jack-of-all-trades' which began as an insult, but now can be thought of more as a complement to well-roundedness of a character. (The second half of the moniker is 'master-of-none', suggesting that the individual is not acceptable focused.) So, it is the case that the main concern will be with the way that the meaning portion of these pairs is altered.

C. Social Lexicon

From this notion of pairs, we see that a group of language speakers have a common understanding. This is the abstract set of word/phrases paired with their meanings. This is not a one-to-one pairing, as discussed, but we can certainly ascertain that many words will have a dominant meaning, or are seen in a context that supports a particular selection from the set of meanings. Establishing that there is a social lexicon-semantic mapping is important because it allows us to take the next step and discuss lexical injection.

D. Lexical Injection and Lexical Exclusion

We will consider lexical injection as the process in which lexicon-semantic pairs are *added* into the social lexicon. Note that this is not adding meaning to words that are already established. Lexical injection consists of a new lexical unit becoming readily understood with the correct semantic value.

This is contrasted with the process of semantic exclusion, where we have new meanings associated with lexical units previously established being changed or altered. This process will describe the ways that new meanings are established for word or phrases that already have semantic values. We will also consider the use of new semantic contexts as being semantic injections. This is to say that new meanings can be attached to a lexical unit either through explicit or implicit means. Now we can consider semantic shift. This is the case where a lexical unit becomes most often associated with a new semantic meaning. When we see the dominant (or most socially acceptable) pairing of lexical-semantic meanings changing, we can say that the particular word or phrase is undergoing semantic shift.

E. Time Series Model

Now that there is a framework for understanding the way a lexicon changes, consider a way to model this process. We will be taking a time series approach. This is a reasonable strategy because we can gather data form social media and examine the trends that are found there. Specifically, the search will be for the ways that we see lexical items in the social media text. A time series is an analytical method that considers some form of signal through time. In this case, we will be looking for the designated lexical items within social media text. This will provide a count, which can suggest insight into the ways that these lexical items are entering the social lexicon.

F. Model Options

In order to find an accurate model, three types will be tested. The first is a logistic growth model. A logistic growth model is a mathematical model used to describe the growth of a population over time. This model is commonly used in ecology, biology, and other fields to model the growth of populations that are subject to resource constraints. The model is based on

the assumption that populations will grow exponentially until they reach a certain size, at which point growth will begin to slow down due to limited resources.

The logistic growth model is characterized by an S-shaped curve, which represents the growth of a population over time. At the beginning of the growth curve, the population grows rapidly, but as the population approaches it carrying capacity (the maximum number of individuals the environment can support), the growth rate begins to slow down. Eventually, the population reaches a point of equilibrium, where the birth rate and death rate are equal, and the population size remains constant. The logistic growth model is useful for understanding the dynamics of populations and predicting future growth patterns, and it has been used to study a wide range of organisms, from bacteria to humans. For changes in the social lexicon, we will be considering whether, once introduced, a specific lexical meaning pair enter the lexicon rapidly, and then maintains an elevated rate as described by the logistic model.

Another model we will attempt to fit is a cyclic model. A cyclic model is a type of theoretical model that describes the object at hand as undergoing a series of cycles or oscillations. These cycles can be thought of as a repeating pattern of increase and decrease, with each cycle being preceded by a period of decrease and followed by a period of increase. The cyclic model is often used in microbiological studies and studies of weather and populations.

If we consider certain predictable events and the ways that they effect social lexicon, we could reason that it will be the case that some of the lexical items enter and exit the social lexicon cyclically.

Finally, we could consider the Hawke's process a reasonable model. Hawkes process is a stochastic point process that models self-exciting events that occur in a sequence, such as

earthquakes, social media posts, or stock trades. The process is named after Alan Hawkes, who first introduced it in his seminal paper in 1971. The main idea behind the Hawkes process is that the occurrence of an event increases the probability of subsequent events occurring in the future.

In a Hawkes process, each event generates a cluster of events around it, creating a cascading effect. The process is often represented graphically as a branching tree, with each event representing a branch point and its associated cluster of events representing the branches. The intensity of the process, or the expected number of events per unit of time, is calculated by summing the base intensity, which is the background rate of events, and the conditional intensity, which is the rate at which events occur given the past history of the process.

The Hawkes process has a wide range of applications in many fields, including finance, seismology, and social media analysis. In finance, the process is used to model the arrival of trades in the market and to estimate the risk associated with a portfolio. In seismology, the process is used to model the occurrence of earthquakes and to predict the likelihood of future earthquakes. In social media analysis, the process is used to model the spread of information on social networks and to identify influential users. The Hawkes process is a powerful tool for understanding complex systems and for making predictions about future events based on past observations. It is his last item that is the most relevant to the study at hand.

G. Cognitive Science Mapping

We are seeking a cognitive science backing to the model and so have to consider a few schools of thought on this. First consider CRUM, which approaches cognitive processes as a computational representation. As the goal is to create a computational model, this fits well. But to emphasize this, consider the process of semantic drift as described above. We can represent lexicon as a set, say a basket, of word meaning pairs. Over time, and with interaction with others

who speak the language, we can have the process of selecting a pair, and adjusting the meaning of the lexical unit.

This would be a strong supporter of the CRUM framework. We can cognitively calibrate the mapping of lexical unit to semantic as time goes on, much like a computer changing representation of inputs. Individuals can also add and subtract semantic/lexical pairs as though we had a dictionary inherent in our linguistic process. This representation fits nicely inside of the CRUM hypothesis.

One way though, that one may view this representation against CRUM is that it is known that social interaction is crucial for semantic shift. Indeed, one's own instance of lexical/semantic pairs are quite crystalline, and were it not for interaction with others, we would not have any force applied to the semantic shift that we see take place. CRUM alone would struggle to account for the random and continuous scale to which another would interpret a lexical semantic relationship that another would devise let alone the thousands of people we will here speak in our lives. And if we were to require a comprehensive model that indicates the probability and to what extent we will come into contact with these near infinite instantiations of 'dictionaries' mentioned previously, we would be pushing CRUM to its limits.

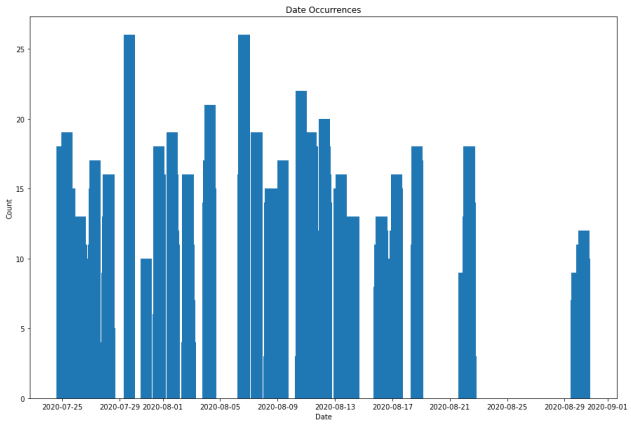
It is for this reason as well that the rationalist school of thought would not be highly supported by this model of social lexicon and the operations that are conducted with/on it. What is written above would tend to refute that there is an innate or inherent meaning to the words. If it is that case that each individual has their own instantiation of lexical-semantic relationships, which, through exposure, pull each other in semantic drift, then it would be difficult to resolve this with the idea that one can, with effort, devise an objectively accurate or even definitive pairing of semantic meaning with a lexical unit.

However, there is a piece of rationalist school that I think is very important to consider here: if we look at the *processes* of semantic change, and all of the operations that can be done on the social lexicon as described above, it would be very comfortable to say that these are inherent. To this end, one could probably satisfactorily argue that the pulling and pushing of meaning (semantic shift) and the addition of words and phrases coupled with the subtraction of others (injection, exiting) is an inseparable aspect of language, and a cognitive underpinning. We know that children can learn, for example, any language they are presented with, and this is done through a very evident linguistic experience. (Though learning a language not knowing any, as infants do is different from changing the meaning of words you are already familiar with.) They do this naturally, so this would be a useful characterization of the processes of semantic change inside of the rationalist school.

Perhaps the connectionist viewpoint has much to offer this approach of understanding semantic shift and social lexicon. If we leverage the idea of distributed systems that are included in the connectionist school of thought, then we can say that by allowing the connections to take place both in a single person's minds, and between the mind of two people, then we can comfortably attribute semantic drift and all the other operations discussed here to interactions between them.

III. Experiment Design

As the goal of this paper is to use a computational model to understand language change, the design was to test these models and evaluate their prediction. In order to test these models, a case study data set was used. This set was procured from https://github.com/gabrielpreda/covid-19-tweets. It is a collection about 17,000 tweets from July-August 2020 with a hashtag ‘covid’. While it was limits in time and quantity, for the scope of proposing computational models, it sufficed. Perspectives on ways to improve these models will be offered later as well as a discussion of difficulties with data. A visualization of the counts of tweets by date is shown here.



Covid is a useful model for studying semantic injection. While it may not be term that arose organically form the way people spoke, and even tough it is of technical origin, we see it rapidly becoming part of the social lexicon, and therefore offer a great opportunity to study it. If we look at the process described previously, we see that pair of ‘covid’ to ‘repository virus’ as a lexical item meaning pair, then it is clear that it was injected into the public sphere, and then underwent linguistic process. While the data set is sparse, wat cannot be computationally captured can be studied empirically.

The attached code provides a more in-depth process of how the models were fitted. Overall, the data was used and fit to the three different forms of models: logistic, cyclical, and Hawkes. The predicted values were compared to the observed values for comparison.

If we find the cyclical model to be the most accurate in predictions, this would suggest repetitive nature to the lexical entering and exiting of lexical meaning pairs. This would suggest a connectionist leaning to the way the semantic shift occurs. This is because with the constant flux of distributed cognition, it would be most appropriate to understand this as a process that takes place outside of an individual.

If the Hawkes process model seems to be the most accurate, we again learn from the connectionist school in the ideas of language change under outside influence, however we would still be left to explain the decay process which would probably come from the individual themselves suggesting perhaps a rationalist viewpoint.

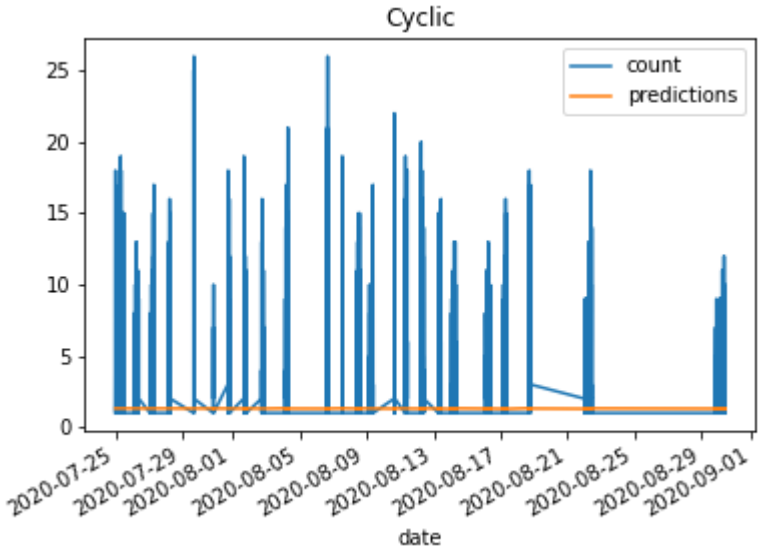
Further, the rationalist view point would be most inline with the logistic model being the most accurate. This is because we see a natural relation between inherent understanding of, and use, of the social lexicon. This use demands that a speaker has a firm control of the meanings and is up to take on all current lexical processes. It is this requirement to function which seems to be innate to human communicators and therefore can be thought of as from within.

There would also be some support for the CRUM model if any of these are a strong fit, given that there is highly computational nature to all of this. If we are able to parametrize these then we can reasonably assert the CRUM is an effective tool for understanding the way that the semantics of a meaning drift.

IV. Results

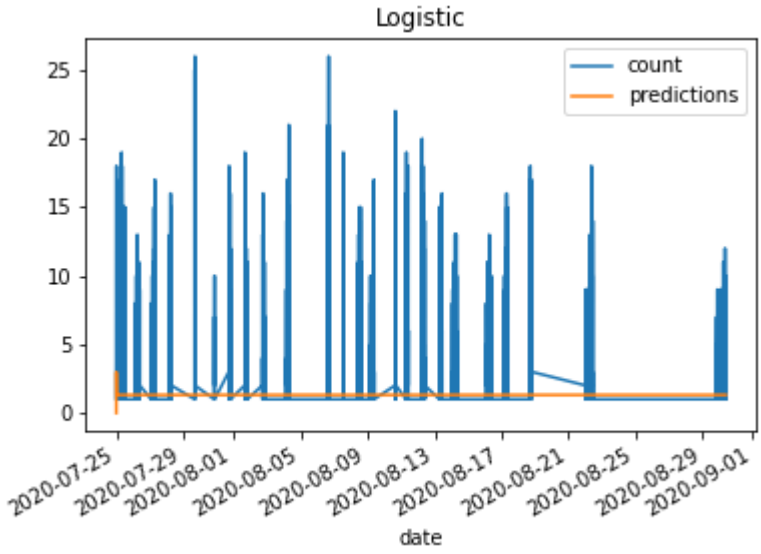
A. Cyclical Model

The plot for the cyclical model is shown here. We see in blue the true values and in orange the predictions. It is quite clear that these was a great discrepancy between the two. While the average prediction of the model appears to be close numerically, at the greater values the model seemed to have a difficult time picking up on ay signals. In the use case both of primary cyclic functions, sin and cosine, were used to model, and a regression was doe to try to learn the linear combination of the two. A discussion of the results will follow this.



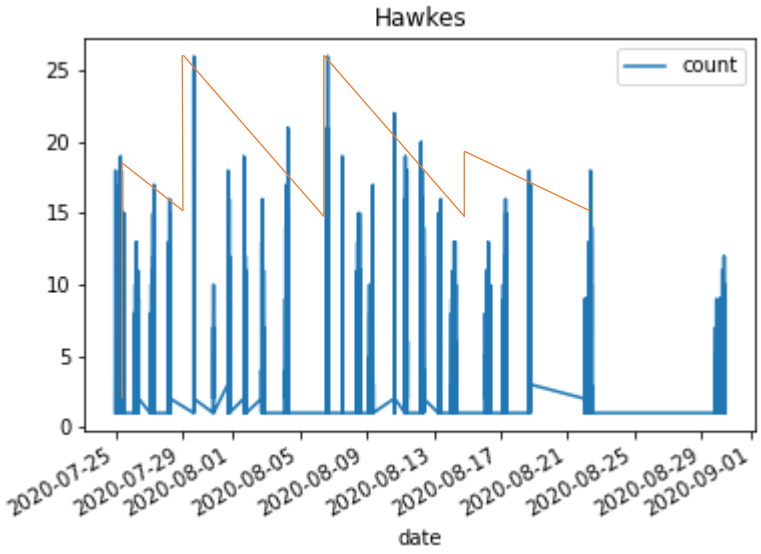
B. Logistic Model

The plot for the logistic model is shown here. It would not be surprising that there was only a small signal at the beginning of this that indicates the model is not accurately predicting the number of times we see 'covid' as a hashtag. The classical S-shaped curve it not presents to the eye in the data, which would suggest that the logistic growth model is not a comfortable predictor of this. The model attempted, it appears, to pick up the group but quickly met its cap in which it became the case the was not a strong enough signal to increase the rate. Again, the true values are shown in blue, and the predicted in orange.



C. Hawke's Process

We see in the case of the Hawkes process that the continuous time series data was not present in this data source. Therefore, computationally fitting this model through programming is a challenge. Consider then an empirical study of the data. There appears to be some slight cyclic process. But it is that case that the cyclic model was not able to strongly pick up on these. There for



woe could consider that the Hawkes process would be a strong contented to replace as shown here, with the proposed markers in orange, and the true values in blue.

V. Discussion

Of course, the data quality and quantity presented a true issue to this study. While computationally there was limited opportunity for the models to pick up on the process, if we consider an empirical analysis to support these then we can discuss the finding. The first is that we do see some cyclic trend in the data. There was a month timespan over which this was collected. We can see that there was some level of increase and decrease in the number of tweets that contain covid. At this point in time, covid had impacted the world for a few months, so one may consider that the lexical injection of the word had already taken place. This would then suggest that the cyclic nature of these tweets could largely be explained by the cyclic nature of tweets in general. As more people have free time on the weekends, more will use that time to engage with social media.

Considering the Hawkes process, we see that the cyclical nature could be explained by this. If we are comfortable suggesting that there is a level of excitement and decay, then we could even say that there is certainly a connectionist idea going on here, where the exciting actions of people on the social media sites connect and distribute this lexical cognition.

There seemed to be a very negligible insight offered by the logistical model. Computationally there was not a clear fit. Empirically we see that the data does not display the characteristic curve. Again, the data was taken after covid had been in the public mind for some time, so we can comfortably say that at this point in the semantic process the lexical item and meaning pair were certainly fully injected in to the social lexicon.

VI. Conclusion

From these results we can ascertain interesting conclusion. Empirical examination will suggest that there is a high degree of connectionist ideology at play. That is, if we view cognition through this lens of distributed cognition, then we can see that semantic drift is a highly connectionist one, with the ability of individuals to communicate being at the forefront of causes of semantic drift. Further we see that there are trends that appear at a macroscopic level that suggest this framework for understanding semantic shift is useful

The CRUM model can also be reviewed under the light here. We see that with limited about of data it can be hard to test. The CRUM model assumes a vast amount of input. The mind contains a debatably, but certainly, high amount of input to perform its cognitions. With out this level of processing ability, true computation methods like the ones developed for this project would not be able to function, as described here. For this reason, CRUM did not seem to have a drastic role in understanding semantic drift, however, were the code able to process the entirety of all of Twitter through the pandemic the expected patter s that were scene in a rough way here would likely be present.

Finally, while only a month of data was shown due to the computational restrictions on hand, it is clear that the are patters and regularities to these sorts of things. The term ‘covid’ would have undergone high levels of lexical injection here. Still, the small cyclic trends that were discussed presented well, which allows us to lend support to the connectionist school.

VII. Limitations and Extensions

Perhaps the clearest limitation to this study was the sparsity of data. It was the intention to pull a much more robust dataset. However, all applications for develop accounts were denied

and there was not the option to purchase large data sets. This severely compromised the computational side of the project. But the code was written to be robust and therefore would, with a largely data set be able to incorporate and fit much more satisfying models.

Further, the data that was put together, was able to be analyzed empirically. This means that there was some insight to be gained. However, to extend this project clearly more data would be beneficial. There are also other forms of models would be useful to test.

Other lexical items would be useful too. Not only hashtags, but analytic the contents of the tweets would help to learn about this semantic shift. We can also expand to other social medial sites which can open the direction for the study to continue with different forms of communication.