Prospectus

Text Classification and Topic Importance Over Time

(Technical Report)

Story-telling Spaces: Fostering Empathy Through Crowdsourcing

(STS Research Paper)

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Harish Chandrasekaran

Spring, 2021

Department of Computer Science

Signed: ___Harish Chandrasekaran_____

Approved: _____ Date _____

Professor Travis Elliott, Department of Science Technology and Society

Approved: _____ Date _____

Ashwin S Kumar, Doctoral Student at Harvard - MIT Division of Health Sciences and Technology

Introduction

As humans have evolved over time, engagement with others in the community has always been a vital part of establishing trust and building connections. While this has not changed, its mode of delivery has taken various forms from pictures to novels and now the internet and smartphones. While many of us take to social media such as Facebook, Twitter, and Instagram to share our stories, I argue that this is not true story-telling. Instead we burden ourselves with how other people perceive us and what it means for our social image. In a vibrant and growing city like Charlottesville and with the technology available to us, we should encourage the sharing of people's stories and backgrounds, especially those that may have been forgotten or erased.

The technical aspect of my paper will focus on using a field of Computer Science called Machine Learning and more specifically text classification and tagging to identify important topics on social media and how those change over time in a certain location. By conducting this analysis, I aim to explore what is important to a community and how we can use this information to effect change. My STS thesis will focus on how storytelling can bring a community together and how we can strengthen the connections between people in Charlottesville by leveraging crowdsourcing technology.

Technical Report

Many of us start our days by turning off our alarms and checking social media on our phones. I personally open Twitter and look at the "Trending" section. However, I have always wondered what factors go into these decisions about what stories and posts should be trending for a specific audience. For my Technical Research, I wanted to look into how topic importance

changes over time in certain areas and how we can identify what is important to a social group based on their social media activity. As mentioned previously, this requires the use of a wide array of Machine Learning and text classification algorithms.

Text classification is the task of assigning tags or categories to text according to its content. It has many applications such as sentiment analysis and spam detection and can be done either manually or automatically. Doing this task manually requires a human to read and produce some terms of importance which summarize the main ideas of the document. Obviously, this proves to be a very tedious and laborious task. And so, Computer Scientists developed a field of research in which computers learn how to complete tasks given some training, namely, Machine Learning (ML). Over time, the processes used to train the computers has become better and computers can now correctly identify hand-written digits with over a 99% accuracy (LeCun et al.). Furthermore, due to these developments in ML algorithms, text classification is quite easy and therefore has become commonplace in everyday applications.

In addition to better text classification, we have seen the rise of recommendation engines, as almost every webpage utilizes them. For example, when you search for a product on Amazon, the webpage also returns a section called “Frequently Bought Together” or “Customers Also Bought These Items”. These recommendations are produced by using pattern recognition and statistical algorithms. Some of these algorithms include K-Nearest Neighbors, Naive Bayes, and Support Vector Machine (Dharmadhikari et al., 2011). K-Nearest Neighbors is a pattern-recognition algorithm which finds the closest matches among the training data to weight possible categories for the document in question. On the other hand, a statistical algorithm like Naive Bayes “makes independent assumptions” by computing probabilities using Bayes

Theorem. While both of these can achieve high levels of success, there are several conditions under which they break down and become more inaccurate (Sehra and Nayyar, 2013).

In the next steps of my research, I will look into deep-learning algorithms, which are computationally much more intensive than the algorithms I have discussed above, but also yield better results. I will also look at how applications such as Twitter and Facebook can determine trending topics not only based on a recommendation basis but also based on a user's location and interests. Using data from a variety of social media, I hope to be able to find out what is important to the local Charlottesville community. I also aim to track these trends over time, as this could provide insights into how the community has grown and evolved. Furthermore, there is potential for this to be applied in other spaces like public policy and governance as this analysis could identify problems in the community and help lawmakers make decisions which uplift the people and community.

STS Thesis

Throughout history, the experiences of underprivileged and underserved communities are constantly forgotten. Without a medium in which to speak, write, or share their stories, the lack of inclusive story-telling spaces has halted our government, leadership, and our society more generally from having the opportunity to learn from the mistakes of the past. In order to remedy this, it is crucial that safe, inclusive, story-telling spaces are formed. In an attempt to form such spaces, my STS group and I have proposed to create a mobile application that allows people in the university and local Charlottesville community to share their personal stories and read those of others. Through the UNST discussions, we were inspired by the untold stories of the enslaved

laborers who built the University of Virginia. The community of enslaved laborers at UVA was crucial for not only the physical construction of the university, but also the day-to-day operations of the school after its founding. Unfortunately, our local community knows very little about the lives of these people as they did not have a medium on which they could share their personal experiences. Furthermore, our discussions in class regarding the forgotten stories of enslaved laborers reminds me of the suppression of stories of Hindu women within ethnic conflicts in India. For example, Hindu and Muslim women in Assam are often left out of the history of India, as the experiences and perspectives of men dominate story-telling spaces, causing there to be a lack of recognition towards the contribution of Indian women in the development of India. However, in today's modern, technologically advanced society, individuals are able to use communicative mobile applications to interact with people vastly different from themselves, allowing society to become more empathetic and understanding. We hope to do the same through our crowd-sourcing, story-telling mobile application.

While the effects of our story-telling application work in theory, I am curious whether true positive outcomes are possible through this application. Thus, I ask the question, "*Can the interaction of technology and inclusive story-telling spaces, such as the mobile application we have proposed, create a more empathetic community that is inclusive of diverse individuals and a growingly international community?*" Based on the research I have conducted, crowdsourcing has proven to be highly successful in a variety of cases and has created a "new culture of openness" which has "changed the way government and the private sector think about engagement and of the role that citizens… in improving innovation (Simperl, 2015). For example, New York University (NYU) was able to use collaborative crowdsourcing to foster a sensitive conversation with 4.500 students and 2,100 faculty members from a wide range of

cultural, religious, and socioeconomic backgrounds from a variety of countries. Furthermore, while NYU's tool was not a mobile application, the online tool was able to form a mechanism that was both anonymous and interactive. It seems that by applying an anonymity feature, individuals in the community felt more comfortable to share their more personal stories, which drastically improved the quality of the conversations (Lorenzo, 2018). It will be crucial for us to include a similar anonymity feature in our mobile story-telling application in order to foster high interaction and story-sharing among users.

In addition to an anonymity feature, it will be vital to include a feature that allows users to not just share their stories, but to read the stories of others and interact with other users. This will foster more understanding within the university and local Charlottesville community. Furthermore, these features will cause new users to continuously use the app, but will also allow new users to be interested in participating within the story-telling space of the application. Furthermore, to achieve true inclusivity and representation of users within the Charlottesville community, it is crucial that the mobile application is free allowing it to be available to individuals with various socioeconomic statuses, that the application is promoted within a diverse group of communities— African-American, Caucasian, Asian-American, Hispanic, etc. Furthermore, the application should also be promoted within the Charlottesville refugee community as the percentage of refugees is increasing annually. Promoting the application in all of these communities is important in order to ensure that there is representation among various diverse groups in the local community. Lastly, including a voice recording feature on the application is crucial in order to create an inclusive space. Doing so will allow non-English speakers to tell their stories, that can then be translated by experts. This allows others in the community to learn from international migrants that do not speak English and learn from their

experiences, causing them to become more empathetic towards the international community. Thus, including all of these various features ensures the success and therefore continuous use of the application.

In terms of our next steps, we will need to engage in community outreach and engagement to receive user feedback. Furthermore, as mentioned above, we will also promote the application in a variety of Charlottesville's local communities— whether that be variation in ethnic background, socioeconomic status, race, younger, or older communities.

Overall, I hope my research in Machine Learning and the research of my group in community engagement can create a stronger and more empathetic community in Charlottesville through story-telling. I also hope to see the diverse cultures and various backgrounds leverage cutting edge technology to solve problems and create a smarter Charlottesville.

References

Dharmadhikari, S. C., Ingle, M., & Kalkarni, P. (2011). Empirical Studies on Machine Learning Based Text Classification Algorithms. *Advanced Computing: An International Journal*, *2*(6), 161–169.

LeCun, Y., Cortes, C., & Burges, C. J. C. (n.d.). THE MNIST DATABASE. Retrieved November 8, 2019, from http://yann.lecun.com/exdb/mnist/.

Lorenzo04/26/18, D. D. (2018, April 26). Using Collaborative Crowdsourcing to Give Voice to Diverse Communities. Retrieved November 8, 2019, from https://campustechnology.com/Articles/2018/04/26/Using-Collaborative-Crowdsourcing-to-Give-Voice-to-Diverse-Communities.aspx?Page=1.

Sehra, S. S., & Nayyar, A. (2013). A Review Paper on Algorithms used for Text Classification. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, *2*(3), 90–99.

Sherman, A. (2011, July 20). How 3 Cities Are Crowdsourcing For Community Revitalization. Retrieved November 8, 2019, from https://mashable.com/2011/07/20/crowdsourcing-city-tech/.

Simperl, E. (2015). How to Use Crowdsourcing Effectively: Guidelines and Examples. *LIBER Quarterly*, *25*(1), 18. doi: 10.18352/lq.9948