

Capstone

Functional Requirements

Sentiment Analysis and NLP

Matthew E. Miller, Manisha Kumari, Lucki Ratsavong
9-26-2019

*** Includes Problem Statement from 9/26 – No Updates ***


Table of Contents

▪ Context Diagram	...pg 2
▪ Use Case Diagram	...pg 3
▪ Data Flow Diagrams	...pg 4-5
▪ Timeline with Collaborators	...pg 6
▪ Interview Questions (From Elicitation Plan) <ul style="list-style-type: none">○ Customer 1 – Ms. Linsey Myers○ Customer 2 – Dr. Khan○ Customer 3 – Dr. Tang	Appendix
▪ Updated Problem Statement (No Updates)	Attached to BB Submit

Instructor Note (**RESOLVED**):

See Appendix

Capstone Lab - Waiting on customer interviews (scheduled 9/23) Inbox x



Matthew Miller <mmill199@kent.edu>
to clenhoff ▾

8:21 PM (2 hours ago) ☆ ↩ ⋮

Caitlyn,

Hope all is well.

My team was only able to schedule customer interviews on 9/23. I am going to add a note of this in our Elicitation Plan. Everything else will be included.

We will submit a revision plan with answered questions on 9/23. We will also have customer interviews (per Dr. Samba).

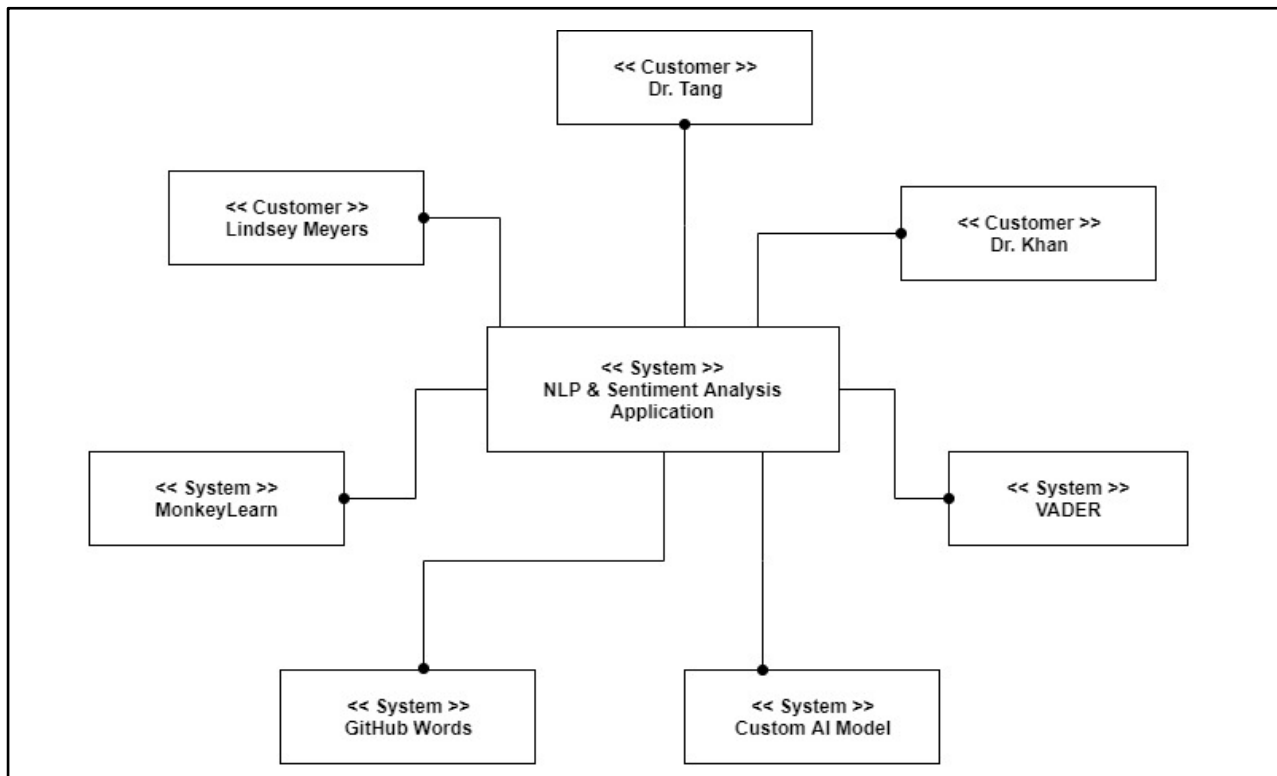
Our interviews are scheduled as follows:

- 9/23 10-10:30a Dr. Khan (CS Dep. Chair)
- 9/23 11-11:30a Ms. Myers (University Marketing)
- 9/23 12-12:30a Dr. Tang (Journalism Dep.)

Kind Regards,
Matthew E. Miller

*** Includes Problem Statement from 9/26 – No Updates ***

Context Diagram:

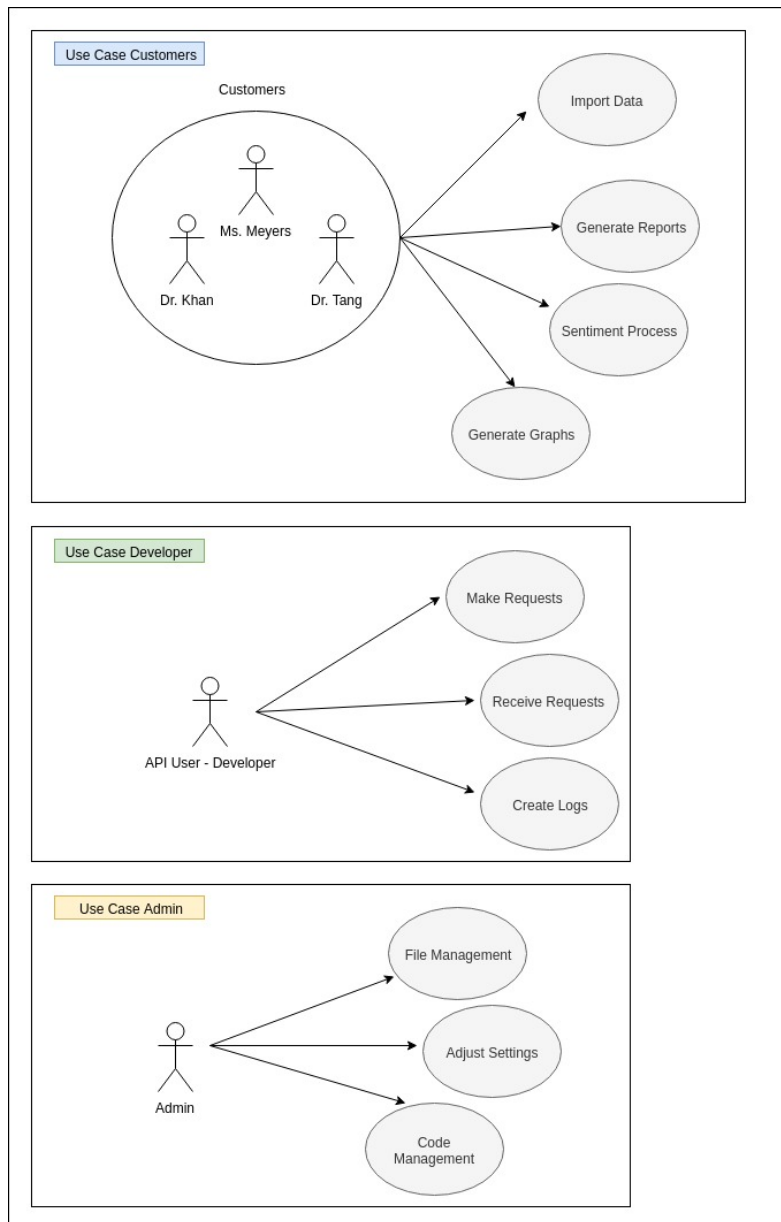


Context Diagram Description:

This context diagram defines the external environments with which our primary system, the NLP and Sentiment Analysis Web App, will interact with. At the center of the diagram is our primary system, the NLP and Sentiment Analysis App. Above it, there are three external systems which represent our customers and their external data. Below them, are four entities that represent the primary external processing systems which give our application its “horsepower”. These four external systems, MonkeyLearn, Github Words, VADER, and Custom AI are the primary drivers of our applications functionality.

*** Includes Problem Statement from 9/26 – No Updates ***

Use Case Diagram:



Use Case Description:

This use case diagram represents the simple interactions that take place between our application and its stakeholders. The first diagram represents our customers and the common ways they intend to use our application. Luckily, our users have similar needs and desires. They mostly want to import data, generate reports, conduct sentiment analysis, and create visualizations like graphs.

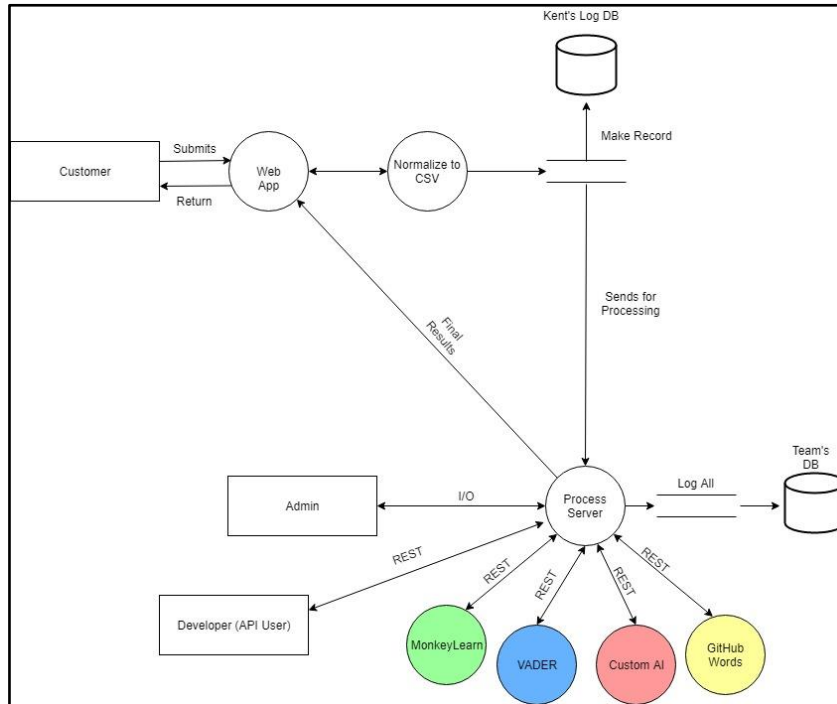
The second diagram represents our API users (developers). They simply want to make requests, receive requests, and create transaction logs. The developer users are just looking for a simple way to interface with our web application and its features.

The third and final diagram represents our administrative users and their common uses. Our system admins want to manage files, configure system settings, and manage code.

*** Includes Problem Statement from 9/26 – No Updates ***

Data Flow Diagrams:

Full System Data Flow:



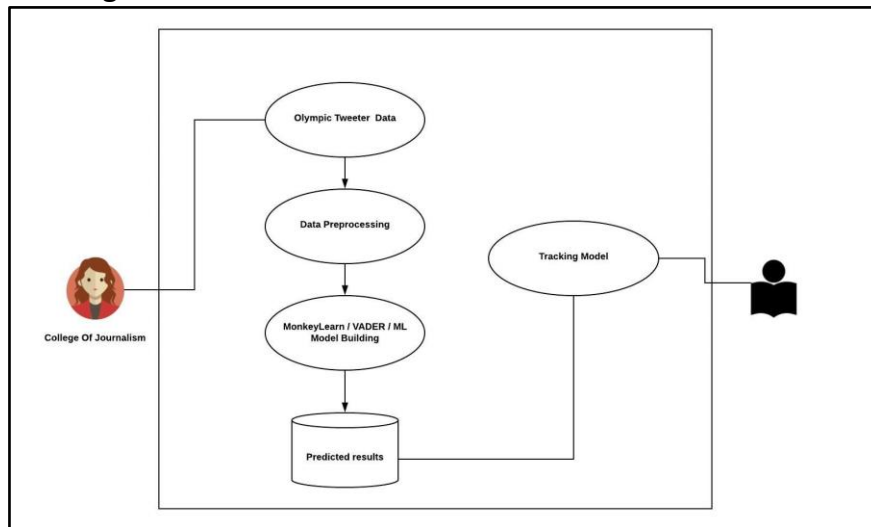
Description:

This dataflow diagram represents the transition of data throughout the NLP and Sentiment Analysis Application.

As you can see, most transactions start with the customer (left) as they import data into the web application.

Once data has reached the web app, it moves to a normalizing process, becomes a transaction record in the database, moves to NLP processing, is DB-logged once more, and then finally returns to the web application as report data.

Dr. Tang Data Flow:



Description:

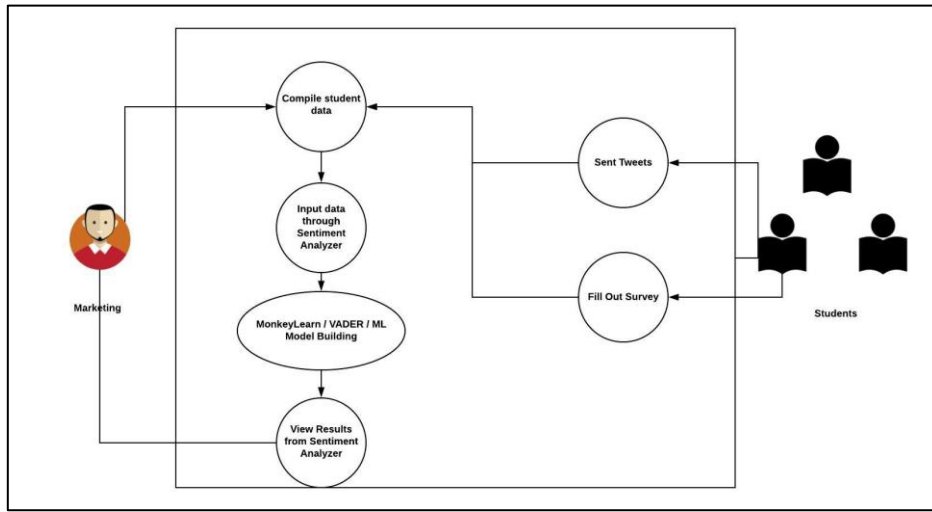
This dataflow diagram represents the flow of data from our customer, Dr. Tang, through the application.

The data comes from twitter, is NLP-processed, results are recorded, and returned to the user.

*** Includes Problem Statement from 9/26 – No Updates ***

Data Flow Diagrams (cont):

Lindsey Meyers:

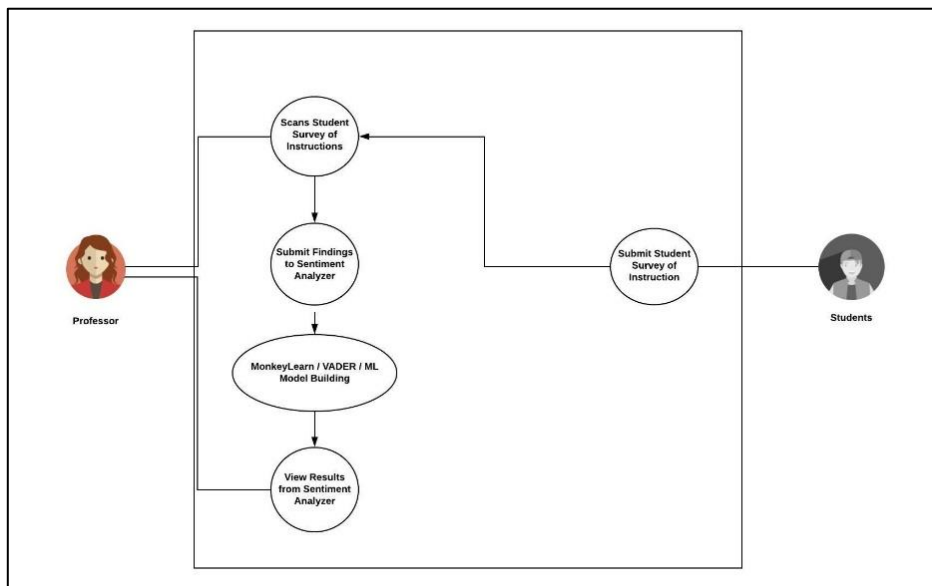


Description:

This dataflow diagram represents the flow of data from Lindsey Meyer's marketing department through the NLP-application.

The data is initially compiled, then placed into NLP-processing, and the final results are displayed.

Dr. Khan Data Flow:



Description:

This dataflow diagram represents the flow of data from Dr. Khan's department through the NLP application.

First, data is scanned to digital text, then placed through the NLP processing tool, and the results are returned.

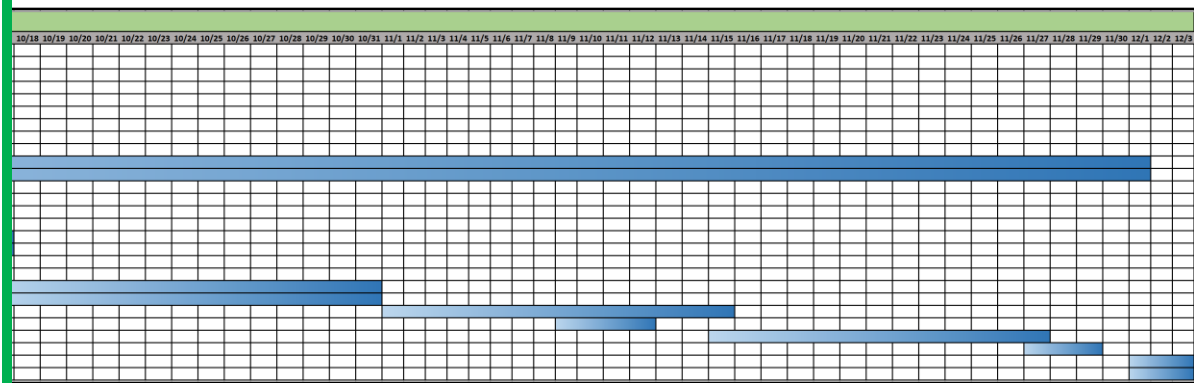
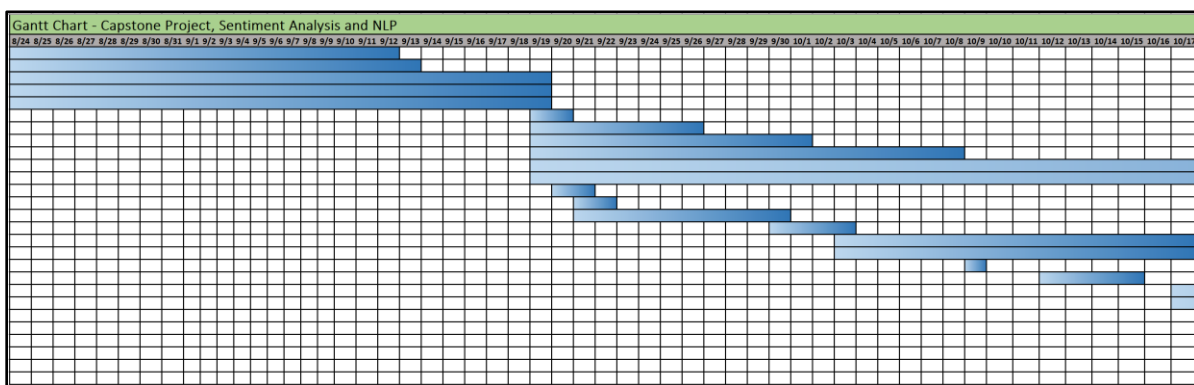
*** Includes Problem Statement from 9/26 – No Updates ***

Timeline With Collaborators:

Schedule						
Task Name	Start Date	End Date	Duration	Percent Complete	Contributors	
First Deliverable - Problem Statement	8/24/2019	9/12/2019	19	100%	Manisha, Matthew, Lucki	
Email and Visit Potential Customers and Users	8/24/2019	9/13/2019	20	100%	Manisha, Matthew, Lucki	
Identify Project Stakeholders	8/24/2019	9/19/2019	26	80%	Manisha, Matthew, Lucki	
Interview Customers and Users	9/12/2019	9/19/2019	7	10%	Manisha, Matthew, Lucki	
Second Deliverable - Elicitation Reqs. and Statement	9/12/2019	9/19/2019	7	20%	Manisha, Matthew, Lucki	
Setup Python Anywhere Server	9/19/2019	9/20/2019	1	100%	Manisha, Matthew, Lucki	
Third Deliverable - Functional Reqs. and Project timeline	9/19/2019	9/26/2019	7	0%	Manisha, Matthew, Lucki	
Fourth Deliverable - Present Progress Report - Systems Architecture and Software Design	9/19/2019	10/1/2019	12	0%	Manisha, Matthew, Lucki	
Fifth Deliverable - Systems Arch. and Software Design via BBLearn	9/19/2019	10/8/2019	19	0%	Manisha, Matthew, Lucki	
Develop NLP and Sentiment Analysis - AI Model	9/19/2019	12/1/2019	73	5%	Manisha, Matthew, Lucki	
Develop NLP and Sentiment Analysis - Syntax/Dictionary Style	9/19/2019	12/1/2019	73	5%	Manisha, Matthew, Lucki	
Setup Project's Dir. Structure	9/20/2019	9/21/2019	1	90%	Manisha, Matthew, Lucki	
Setup Remote Request Processing server	9/21/2019	9/22/2019	1	0%	Manisha, Matthew, Lucki	
Design ER Diagrams	9/21/2019	9/30/2019	9	60%	Manisha, Matthew, Lucki	
Setup Remote Databases - myDB	9/30/2019	10/3/2019	3	0%	Manisha, Matthew, Lucki	
Design Processing Server's Program logic	10/3/2019	10/17/2019	14	0%	Manisha, Matthew, Lucki	
Design User Interface - Web Application	10/3/2019	10/17/2019	14	0%	Manisha, Matthew, Lucki	
Sixth Deliverable - Mid-semester Peer Evaluations via BBLearn	10/9/2019	10/9/2019	0	0%	Manisha, Matthew, Lucki	
Seventh Deliverable - Present Progress Report	10/12/2019	10/15/2019	3	0%	Manisha, Matthew, Lucki	
Implement Processing Server's Design	10/17/2019	10/31/2019	14	0%	Manisha, Matthew, Lucki	
Implement User Interface	10/17/2019	10/31/2019	14	0%	Manisha, Matthew, Lucki	
Connect System Components	11/1/2019	11/15/2019	14	0%	Manisha, Matthew, Lucki	
Eighth Deliverable - Present Progress Report	11/9/2019	11/12/2019	3	0%	Manisha, Matthew, Lucki	
Test, Revise, Maintain	11/15/2019	11/27/2019	12	0%	Manisha, Matthew, Lucki	
Curate Supporting Docs.	11/27/2019	11/29/2019	2	0%	Manisha, Matthew, Lucki	
Final Deliverable - Group Presentations	12/1/2019	12/3/2019	2	0%	Manisha, Matthew, Lucki	
Create Project Presentation	12/1/2019	12/3/2019	2	0%	Manisha, Matthew, Lucki	

Description: The timeline and gantt chart are pretty straight forward. All team members are allocated to each task because our team is so small. The timeline tracks the submission deadlines from the course syllabi.

Gantt Chart (line up with timeline-table):



Appendix

Anticipated Interview Questions:

Lindsey M Myers, M.A., Director, Marketing Strategy and Research (Destination Kent State)

High Level Understanding:

1. Could you please tell us more about your role?
2. What types of data do you use to make decisions?
3. How impactful is this data? Is it the primary driver of decisions?
4. Are they used to determine marketing efforts and campaigns?
5. How do you analyze this data? Current limitations?
6. What could be improved about the current analysis of this data? Likes and dislikes.
7. What are your thoughts about sentiment analysis?
8. Have you previously encountered obstacles with sentiment analysis?
9. How do you think sentiment analysis could benefit you in your role and your department?

Detail Oriented:

1. For both input and output, what should be the format of the data?
2. Must any data be retained for any period of time?
3. Are there constraints on size of the system (Handheld/Server/PC etc)?
4. Are there any COTS or other constraints on programming language, OS because of existing software components?
5. Is input coming from one or more other systems (“upstream”)?
6. Is output going to one or more other systems (“downstream”)?
7. What is the protocol for the upstream and downstream systems?
8. Who will use the system?
9. Will there be several types of users?
10. What is the skill level of each user?
11. What kind of training will be required for each type of user?
12. How easy should it be for a user to understand and use the system?
13. How much data will flow through the system?
14. How often will data be received or sent?
15. When and in what ways might the system be changed in the future?
16. How easy should it be to add features to the system?

The interview questions that I have selected for Lindsey Myers are with the goals in mind of understanding her role, how she utilizes data in her role, current efforts, improvements, and how our project could improve the overall process.

Next Page, Customer 2...

Dr. Khan, Computer Science Department Chair (Course Evaluations)

High Level Understanding:

1. How significant or how much weight do student survey of instruction (SSI) carry?
2. What is the impact of SSI?
3. Do they inform decisions on promotions or salary increase? Whether a course is offered more or less in the future? If a Professor is removed from a course or not? Gaining tenure?
4. Informing instructors on their teaching skills?
5. How long does it take from when the SSI are submitted to what results are given?
6. What does the current analysis of SSI data look like? How could it be improved?
7. What differs from the new online SSI versus the paper ones? Limitations of the current process?

Detail Oriented:

1. For both input and output, what should be the format of the data?
2. Must any data be retained for any period of time?
3. Are there constraints on size of the system (Handheld/Server/PC etc)?
4. Are there any COTS or other constraints on programming language, OS because of existing software components?
5. Is input coming from one or more other systems (“upstream”)?
6. Is output going to one or more other systems (“downstream”)?
7. What is the protocol for the upstream and downstream systems?
8. Who will use the system?
9. Will there be several types of users?
10. What is the skill level of each user?
11. What kind of training will be required for each type of user?
12. How easy should it be for a user to understand and use the system?
13. How much data will flow through the system?
14. How often will data be received or sent?
15. When and in what ways might the system be changed in the future?
16. How easy should it be to add features to the system?

The interview questions that I have selected for Dr. Khan are with the goals in mind of understanding the significance of SSI, current efforts, improvements, and how our project could improve the overall process.

Next Page, Customer 3...

Dr. Tang, Professor - School of Journalism and Design

High Level Understanding:

1. How are the Tweets selected?
2. Do they have a specific hashtag?
3. Include a specific user?
4. Posted on a particular page?
5. What does the process of current analysis look like (i.e. manually)?
6. How long does it take to complete?
7. What results does it offer?
8. Tools used?
9. Limitations?
10. How would sentiment analysis improve upon this process?

Detail Oriented:

1. For both input and output, what should be the format of the data?
2. Must any data be retained for any period of time?
3. Are there constraints on size of the system (Handheld/Server/PC etc)?
4. Are there any COTS or other constraints on programming language, OS because of existing software components?
5. Is input coming from one or more other systems ("upstream")?
6. Is output going to one or more other systems ("downstream")?
7. What is the protocol for the upstream and downstream systems?
8. Who will use the system?
9. Will there be several types of users?
10. What is the skill level of each user?
11. What kind of training will be required for each type of user?
12. How easy should it be for a user to understand and use the system?
13. How much data will flow through the system?
14. How often will data be received or sent?
15. When and in what ways might the system be changed in the future?
16. How easy should it be to add features to the system?

The interview questions that I have selected for Dr. Tang are with the goals in mind of understanding the business value resulting from sentiment analysis, journalism school has a client that wants to know acceptance of this e-sport event (impression) a bunch of tweets, go through and do sentiment analysis 10% manual paper for her client started with Dr. Guan, work with him, or pass work onto us.

Actual interview transcripts on next page...

Actual Interview Transcripts:

Lindsey M Myers, M.A., Director, Marketing Strategy and Research (Destination Kent State)

Tell me about your role?

Director of Marketing Strategy and Research at Kent State.

Before she would have clients from Departments and Colleges and do marketing for them, but once a colleague, who specialized in research left, she decided to take on that role. Today, research takes about 90% of her time.

What data do you utilize to inform your decisions?

Focus group and surveys, where they can analyze student perspectives on various events.

These range from expectations, interest, thoughts, what should be added or removed, improved on.

From this research she will be able to present to clients what talking points to push, features to use, a roadmap for the project.

Interpret and review.

Web team uses usability testing, heat mapping, test designs with appropriate groups, Google Analytics.

What is the impact of this data? Or how significant is this data? What is it used for? Are they used to determine marketing efforts and campaigns?

Always wish that all marketing campaigns were research based, but unfortunately, due to time, based on client needs, not all marketing campaigns are research based. However, campaigns with a lot of funding typically have a solid research backing.

How do you analyze this data?

Open ended tally of positive, neutral, negative comments. Manual.

Group responded by content to hopefully find an actionable solution.

Context clues.

Would love to catch early, so they can improve as they go!

Current limitations?

****TIME!!**** Also, for manual, this is typically handled by student workers, so the coder reliability fluctuates. SurveyMonkey beta sentiment analysis was not helpful, produced inaccurate results. If you can produce a 95% satisfactory rating that is amazing and good enough to publish in an academic journal.

*** Includes Problem Statement from 9/26 – No Updates ***

What could be improved about the current analysis of this data? Likes and dislikes.

Our product would be faster. They are just about to finish the Destination Kent State project, even though, those surveys wrapped up a couple months ago. Questions that are open ended are a drag and takes a length process.

Student worker looks through the survey.

Download data in any format.

(-) Some print a document and highlight

(-) Some use Excel and put into columns

Either way must do this process of reading the data more than once to develop accurate columns of data.

How do you feel about sentiment analysis? What are the obstacles to employing this?

Valuable!

Final Notes:

- CSV formatting is good for research (maybe incorporate a how-to on website)

Dr. Khan, Computer Science Department Chair (Stand-in Dr. Lu, Asst. Chair)

How significant or how much weight do SSI carry?

This varies from department to department, and per Chairman. Most people are looking at the numbers they receive on SSI and not typically the comments. If you see a number out of range, then will potentially look into the comments. Administration wise - not sure what they do with the SSI data. SSI offer Chairs the ability to observe and see how the new faculty is doing. The SSI may also assist Chairs with writing evaluations on the instructors and also allow them to give recommendations on how to improve teaching.

What is the impact of SSI?

Negative course evaluations - Chair talks to instructor, try to help instructor partner with a senior faculty member to mentor.

Do they inform whether a course is offered more or less?

Yes, if there is are many positive SSI on a course, they will consider incorporating that more into the curriculum and potentially offer it more often.

How long does it take from when the SSI are submitted to when results are given?

For Fall SSI, they receive that at the beginning of the 2nd semester, early February (dependence on the Secretary's schedule). They receive a PDF scan and statistics.

What does the current analysis of SSI data look like? How can it be improved?

They receive a PDF and Excel sheet with quantitative statistics. Improvements - not sure.

What differs from the new online SSI versus the paper ones? Limitations?

Too early to call, have not dealt with that yet. I'm sur that it'll make life easier. This may allow more time to draw conclusions.

Final Notes:

- Roughly 1000 SSI for the CS Department.
- If we built our product, states that it will be helpful to large, entry level courses (freshman, sophomore) required courses.
- Most look into the extreme cases.

Dr. Tang, Professor - School of Journalism and Design

Role?

Professor and Graduate Coordinator for Journalism and Digital Science. Her research focuses on sport communication, in regards to the International Olympic Committee and the E-Sport community. She is interested in studying E-Sport as she is intrigued of the mix of active and passive roles. As a Graduate Coordinator, she is tasked with student recruitment and scheduling.

Where is the data from?

Survey content analysis; secondary analysis; Twitter data; focus group on occasion

How impactful is the data? Is it the primary driver of decisions?

It is impactful, but the primary driver of decisions comes from the analysis that she discovers in her literature review.

How are Tweets selected?

Primarily through keywords. There are two groups of keywords and in one group there is at least "e-sport" and the other group there is at least "Olympics". We collect everything from those who use specific keywords.

Do they have a specific hashtag?

Yes we can search by hashtags, but this produces a limited amount of results.

What does the process of current analysis look like?

Using SPSS to analyze and sift through large datasets. Then performs statistical analysis such as regression test to answer research questions and is then able to describe findings.

How long does it take to complete?

Real quick, less than a week for the analysis. The data cleaning takes a while.

*** Includes Problem Statement from 9/26 – No Updates ***

Tools used?

SPSS for regression and correlation. However, the problem with SPSS is that it can handle only a certain amount of data, unable to do large datasets. She states the limit she believes is 10,000; however the largest she has worked on is 2000 cases and 200 verbals. SPSS is not capable for big data analysis, nor does it work with sentiment analysis.

Final Notes:

- Preferred format: CSV, SPV
- We are going to receive not a large amount of data.
- She is interested in charts, if time permits.
- She is interested in context analysis, in addition to sentiment analysis.