Research Proposal Transcript

Slide 1: Title Page

Hi, and welcome to this brief presentation regarding a research proposal, it's background, significance, related works, questions, and methodology. Given the importance of security on software development, this research proposal focuses on secure software development, specifically, how impactful are security guides on the development community in 2022.

MOVE TO THE NEXT SLIDE.

Slide 2: Background

Developing secure software is no easy task. As we observe, the rate of vulnerabilities has been on an upward trend since 2015 as shown by WPScan Statistics (2022). This can be attributed to the ever-increasing complexity of information systems and an increase in the number of lines of code written, or it could be due to a lack of security knowledge within the developer community.

Regardless, security concerns are not isolated but very much social, given that programming software development is considered a social activity as stated by Stefik and Henenberg (2014). Suppose we are to investigate the impact of security guidance on a developer community. In that case, how would we go about to obtain relevant market data? Fortunately, there is no better source than Stack Overflow, with over 15M users, 21M questions and 32M answers available freely to researchers.

But how do we analyse the Q&A data looking for security insights?

Well, one approach is to leverage topic modelling which several researchers have used. This technique allows researchers to search a large corpus of text and automatically classify each sentence into several unique topics. These topics are then given various distributions which researchers can then use in statistical analysis. Very useful to uncover recurring development issues, according to Johri and Bansal, 2018.

MOVE TO THE NEXT SLIDE.

Slide 3: Significance

Why then is this research relevant for a research project, and what is its significance?

Firstly. Developers often ask closely related questions about the same security topic. For instance, in the example shown, two users have asked a similar question regarding HTTPS security. Looking at the answers, we find no evidence of security guidance as published by security standards bodies, leaving the reader to simply accept answers as true and reliable.

As a result, we ask, "*How concerned are developers about secure development?", "Should organisations mandate security training for their developers?".* Or for the purposes of this proposal, "*What is the impact of security standards bodies on real-world developers in their Q/A interactions?*"

We searched through literature and identified that several researchers (listed in the next slide) have also used the free SO datasets for various analyses to answer their unique research questions. However, as far as we know, no recent research has been undertaken that considers the influence of security standards on Q/A sites.

This research, therefore, is significant because it aims to uncover a correlation between the security standards and development questions and answers. It is relevant because it looks to provide insights within the past two years.

MOVE TO THE NEXT SLIDE.

Slide 4: Related Work

Looking at the literature for this topic, we identified several categories covered by researchers who also investigated the SO data. We also note that most researchers also use topic modelling techniques over the SO datasets.

Similar works related to this research proposal are as follows. Barua et al. (2014) recommends researchers carry out further research of topics hidden in both questions and answers; Lopez et al. (2018) identify the social aspect of developers and how they have *some* awareness of security, which is good; Tahaei et al. (2020) recently looked at the interesting topic of privacy polices and what are the drivers forcing behind developers adopting them. Their results showed organisations like Samsung, Apple and Microsoft are the major drivers behind why developers implement privacy policies.

Interestingly too, Neuhaus and Zimmermann (2010) investigated the CVE vulnerability database but, unfortunately did not link the CVE to SO data and also did not use topic modelling.

MOVE TO THE NEXT SLIDE.

Slide 5: Research Gap

From the research literature available, we therefore consider there is a research gap available for this proposal. Namely, that there is a need to carry out recent up-to-date research to establish the connection between security standards and their evidence inside developer Q/A sites, like Stack Overflow.

MOVE TO THE NEXT SLIDE.

Slide 6: Research Questions

Following from the identified research gap, we now move on to frame the project proposal by asking three questions:

- One. “How important is security guidance as evidenced by a Q/A site?”
- Two. “How often is security guidance referred to in either questions or answers on a developer Q/A site?”
- Three. “How has the impact of security guidance changed as reflected in developers’ Q/A over recent years?”

From the questions, our research proposal now considers the impact of the OWASP Top 10 (2021) security guidelines in relation to a Q/A site. The objectives of the project are broken down into three, each related to a research question:

- One. To obtaining a ranked list of topics
- Two. Relates to generate a corpus of topics from security guidance
- And lastly. Relates to the statistical analysis and correlation of the two datasets that will be analysed and documented.

MOVE TO THE NEXT SLIDE.

Slide 7: Research Methodology

To carry out the project, we propose the following research methodology as listed in tabular format on this slide.

For Approach. We will conduct the project in a quantitative manner.

For Framework. We follow a six-stepped framework briefly described:

- In step 1, prepare and communicate the selected/assigned supervisor, prepare hardware and software environment.
- In step 2, locate required data sets from SO and corpus text from OWASP Top 10:2021.
- Step 3 (especially applicable to the SO data), we will need to pre-process the XML data to strip out irrelevant attributes and potentially manually assign textual categories.
- Step 4, generate topic models from SO data and OWASP Top 10:2021 using relevant plugins
- Step 5, recursively iterate over analysis and tweaking of parameters
- Step 6, finally, gather our analysis results, document them, and present them

For Collection. We will obtain our literature from reputable sources such as Scopus, ScienceDirect or IEEE Explore. All datasets for SO will be obtained from the publicly available Stack Exchange Data Dump.

For Analysis. We will leverage topic modelling, we will either use a Java or Python library as shown. However, this depends on which environment provides the easiest manner to process and analyse the data.

Due to familiarity with Excel, this tool will be used as far as possible for tables and graphs. However, we may leverage other tools provided there is no cost attached.

For Scope. The project will be limited to only contain post that contain the phrase or semantically similar words, “security”. We may need to use stemming or lemmatisation beforehand. Also, our data scope (due to time constraints) will initially be limited to posts within the last two years.

For Limitations. And finally, for limitations to this project, we may come across literature that is hidden behind a paywall.

MOVE TO THE NEXT SLIDE.

Slide 8: Ethical Considerations

Following on, there are not many ethical factors to consider. However, most importantly we will ensure that all personally identifiable information in any dataset is discarded. We will ensure we have the right software licenses for use and any additional tools that we leverage must be free or open source. Lastly academic references will be used wherever other authors’ works have been used.

MOVE TO THE NEXT SLIDE.

Slide 9: Project Timeline

Next, onto the project timeline.

We tentatively propose the following timeline that points to the project taking out 162-man days-including weekends. We tried to keep the number realistically low to allow some buffer room to accommodate for unseen circumstances or delays. Therefore, this value must be seen as tentative.

Most of the project, we think will be spent around the processing and generating of topic models. This is a new area, one which we do not have prior experience, so this part of the timeline may likely shrink or grow substantially and is listed as a risk in the next slide.

MOVE TO THE NEXT SLIDE.

Slide 10: Risk Assessment

Talking of risks, here we consider several points that may harm the project delivery. The primary risk being time, especially due to fulltime work constraints. Second to that, are potential costs to purchase of software for processing capability. However, one mitigation is the potential option to use a free cloud provider trial. But, trials typically only last for 30 days so this may introduce its own issues.

The next risk is lack of experience in generating topic models. However, we looked at various code samples and descriptions and feel that, on the surface, it may not be much of a concern.

The remaining risks are not particularly troublesome, though they are listed here for completeness. For example, the recency of data within the SO dataset might impact the third research question that looks at trends over time. However, we will mitigate this by ensuring that our datasets contain only data that falls within the time period defined for the scope.

MOVE TO THE NEXT SLIDE.

Slide 11: Budget

In terms of budgeting, we have listed basic software and hardware items to be used in the project. However, as listed in the risks above, there might be a need to purchase additional software if this existing list of software requirements is insufficient. However, overall, there are no budget concerns at this point in time.

MOVE TO THE NEXT SLIDE.

Slide 12: Artefacts

Some of the artefacts we envision for the project will include tables, graphs, and data sets. Optionally, we may also provide the configuration values used for our LDA (Latent Dirichlet Allocation) parameters, or other configuration files required for IDE plugins.

Our tables will hold tabular nominal or ordinal data. While the graphs will show visually the relations between topics, or security concerns. We will also produce an analysis and interpretation of the correlations uncovered by the data.

MOVE TO THE NEXT SLIDE.

Slide 13: Thank you.

Finally. Thank you very much for your time in listening to this presentation and allowing us to present and briefly describe each aspect of this proposal. We trust the data meets expectations and look forward to contributing toward this interesting research area in the future.

Thank you and keep well.