Response document

This is a response document for resolving defense comments of master thesis <A comprehensive topic-model based hybrid sentiment analysis system>. There are 45 comments recorded during the defense in total which are broken down to different Chapters in the thesis. The comments for each Chapters are listed below:

Comments Distribution Over Chapters	
Chapter Name	**Number of Comments**
Abstract	3
Chapter 1: Background and Introduction	17
Chapter 2: Literature Review	5
Chapter 3: Proposed Data Acquisition Method	3
Chapter 4: Proposed Solution for Sentiment Classification	4
Chapter 5: Sentiment Classification Evaluation	3
Chapter 6: A Use Case and Pipeline Discussion	5
Chapter 7: Conclusion and Future Work	1
Others	4
Overall number of comments	45

First of all, we want to mention that we keep the original comment numbers without re-index them for each chapter. Therefore, the question won't be shown in a sequential way but will follow the structure of our thesis. Secondly, the grouping of comments doesn't mean only the related paragraphs within that chapter are revised; The revision can happen across chapters and the grouping normally means where the majority revision happened or where the revision starts from. For example, there are comments that are expecting the contribution part to be revised. We categorized it as changes required in Chapter 1 where we explicitly claim our contributions, but any sections mentioning contribution are also revised due to these comments. In addition, some comments that not belongs to any chapter specifically (like "Avoid too many subsections where possible") is put under "Others".

In the following pages, we will respond the comments one by one. For each question, we will describe what the related word/sentence/paragraph/chapter looked like previously, and what change we made during revision. Furthermore, for most of the comments, a table will be used to show the position of the revised part before and after the change, and the corresponding paragraph/text before and after revision which are colored as blue. Some clarifications about the changes are written in *blue Italic*.

Abstract

7. Abstract does not say what the problem is ...

Before revision, the problem statement is vague in the Abstract. I only mentioned there were problems from different part of the pipeline, but I did not specify what exactly are the problems that to be solved.

After revision, I rewrote the abstract, by pointing out what exactly are the three key problems for data collection (the trend that publicly available training datasets are becoming less available), topic-model based sentiment classification (the difficulty to determine topic numbers for topic model-based approach), and the problem about using the proposed models for real business problems in a daily basis (a lack of data level discussion about how to utilize the proposed models day-to-day to drive applications). After the revision, it can help reader to have a better idea what are the problems we want to solve and what are the solutions we provided as responses in a high level (Due to the length limitation of Abstract).

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
Abstract, page ii	Abstract, page ii	For the past decades, sentiment analysis is one of the most the popular topics in Natural Language Processing (NLP). Twitter sentiment analysis is a special subtopic in sentiment analysis task that draws researchers' attention since its potential to drive commercial insights and the difficulty to process tweets compared with blog and newspaper text. We are inspired to propose a complete ecosystem, including data collection, sentiment analysis, and data visualization, to try to solve some remaining problems based on the current studies. By proposing this system, we firstly offer a new way for big volume Twitter training data collection; Also, we proposed a hybrid sentiment classification model on top of topic modelling by using our self-collected tweets corpus, which achieves a 78.54 F-score; Last but not least, our visualization system is the first one considering tracing sentiment trend over time compared with other online tools and systems.	Nowadays, Twitter sentiment analysis is drawing a lot of attention due to its potential to drive decision making in a variety of domains. However, the trend that publicly available training datasets are becoming less available, the difficulty in determining topic numbers for topic model-based approach, and a lack of data level discussion about how to utilize the proposed models day-to-day to drive applications are still the remained concerns. To solve these problems, we firstly offer a new method to collect and build Twitter training dataset based on noisy labels; In addition, we proposed a topic-model based hybrid sentiment classification model by using our self-collected tweets, which utilizes three different topic models and coherence score to choose the best topic model in an automated way; Last but not least, a use case is illustrated to show how

			to apply our pipeline in a daily basis to solve real business problems.

6. Abstract - what is Ecosystem? Better called it as Full pipeline ...

Before revision, we were using “ecosystem” which is not an accurate word to describe the work we have done and can lead to some confusion.

After revision, we replaced the word “ecosystem” with “pipeline”, from the related paragraph that this word appears. The three paragraphs (Abstract, Chapter 1 Introduction, and Chapter 1.2 Thesis Overview) that contained this word before have all been revised now.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
Abstract page ii	Abstract page ii	ecosystem	pipeline
Introduction page 3	Introduction Page 13	ecosystem	pipeline
Chapter 1.2 Thesis Overview page 7	Chapter 1.5 Thesis Overview Page 21	ecosystem	pipeline

9. Vague terms like big volume ... how much is big? Clarify that

Before revision, we used “big volume” in some statements but without clarifying how big could be regarded as “big volume”. It was not clear and may lead to confusion.

After revision, we find all the paragraph that ever mentioned “big volume” previously and revise them to a more appropriate statement.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
Abstract, page ii	Abstract, page ii	big volume	… (We rewrite the sentence and drop the “big volume” phrase) …
2.1.1 Datasets review Page 11	2.1.1 Existing Twitter Datasets review Page 24	big volume	around 40000
2.1.1 Datasets review Page 11	2.1.1 Existing Twitter Datasets review Page 25	big volume	about 100 million

Background and Introduction

17. Opinion mining and sentiment analysis, emotion, terms ... clarify these, how they are different, how these are related? and add details which term(s) you really are using in your system?
For example, in sentiment analysis, some studies ... interchangeably ... clarify that ... which term you are using, and keep it consistent across all the thesis, define it and then use it in the thesis

Before revision, we mentioned some definitions of sentiment analysis from previous works, but we didn't specifically clarify what definition/term we were using in our thesis.

After revision, we add our own definition of sentiment analysis after reviewing several existing definitions. Our definition defines the task and scope of sentiment analysis for our thesis specifically.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
1.1.2 Sentiment analysis, opinion mining, and subjectivity analysis	1.2 Terminology used	-	Based on all these definitions from previous study, we want to define the task and scope of sentiment analysis in our study as following: "Sentiment analysis aims at extracting the sentiment orientation for the given text and assigning them to either positive or negative class. It also includes the techniques that could be potentially used to achieve this goal." By using this definition, firstly we want to distinguish sentiment analysis as a different task from opinion mining and subjectivity analysis. In our thesis, sentiment analysis focuses on assigning positive or negative polarity to the given text, while opinion mining and subjectivity analysis focus more on detecting subjectivity from the text. Secondly, we want to clarify that based on our definition, sentiment analysis is a binary classification task rather than detecting the sentiment strength (one possible way is to assign a score from 1 to 5 based on the strength of the sentiment, or other score scopes, for both positive and negative). We do not consider detecting sentiment strength because of the following reasons: In order to detect the sentiment strength for the given text, training data with sentiment strength is required to contain strength as well which is hard to acquire in an automated way; An alternative could be using lexicon that has

			sentiment strength for each word, but lexicon-based approach performs poorly for Twitter sentiment analysis. Meanwhile, we want to compare our proposed method with previous works, and previous works mostly consider sentiment analysis as a 2-way or 3-way classification task as well. Overall, due to the goal of building an automated pipeline, and making comparisons with previous works, we want to keep our task as a binary classification problem.

14. Topic modeling ... clarify this term

Before revision, we mentioned topic modelling/topic model a lot of times before bringing the definition/context first.

After revision, we add "Topic model or topic modelling" under terminology used section so that the readers can have a better context when reading the thesis. Also, in Section 4.3, there are explanations of the goal of topic modelling and in-depth descriptions of four popular topic models before we go into details of our proposed topic model based sentiment classification model.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
-	1.2 Terminology used	-	**Topic model or topic modeling.** Topic modeling refers to the task that aims at uncovering the latent topics from a collection of texts which is firstly proposed in [5]. While topic model refers to the specific model that help to achieve this goal, like Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and so on. There will be detailed explanations in Chapter 4 before we have an in-depth discussion of applying topic modeling in our approaches.

11. Clarify, introduction (page 2), hybrid system, clarify what exactly do you mean by hybrid ... it is unclear ...

Before revision, we didn't explicitly define what does "hybrid" means in our context which may lead to some confusion.

After revision, we add "hybrid or hybrid approach" under section 1.2 terminology used so that the readers can have a better understanding what we are referring to when saying "hybrid" or "hybrid approaches" in our thesis.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	1.2 Terminology used	-	**Hybrid or hybrid approach.** Generally, the word "hybrid" or the phrase "hybrid approach" is from the previous works of Twitter sentiment analysis which refers to the method that combining two or more existing approaches together to build a new approach. The approaches that being combined could be from lexicon-based approach, the subcategories of machine learning approaches (supervised learning based approaches, semi-supervised learning based approaches, and unsupervised learning based approaches) but not limited to them. In our thesis, when we are discussing our proposed sentiment classification model, we use the "hybrid" to refer to our method that combines topic modeling (unsupervised learning), clustering (unsupervised learning), ensemble learning and single classifier building (supervised learning) together. In the literature review section, "hybrid" is used as the general meaning.

23. In introduction, define all terms that you use in your thesis

Before revision, we used some terms to express some specific meanings without explicitly define them under our thesis's context which could lead to some confusion.

After revision, we define the terminology used in our thesis, either to clarify some existing terms, or define some specific terms in the context of our thesis. Also, we restructure the related section to make it more concise and consistent by creating a new section 1.2 Terminology used.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
1.1.1 Sentiment, motion, and opinion 1.1.2 Sentiment analysis, opinion mining, and subjectivity analysis	1.2 Terminology used	…Definitions about Sentiment, motion, and opinion & Sentiment analysis, opinion mining, and subjectivity	…Definitions about Sentiment, motion, and opinion & Sentiment analysis, opinion mining, and subjectivity analysis in 1.2… **Hybrid or hybrid approach**. Generally, the word "hybrid" or the phrase "hybrid approach" is from the previous works of Twitter sentiment analysis which refers to the method that combining two or more existing approaches together to build a new approach. The approaches that being combined could be

		analysis in 1.1.1 and 1.1.2…	from lexicon-based approach, the subcategories of machine learning approaches (supervised learning based approaches, semi-supervised learning based approaches, and unsupervised learning based approaches) but not limited to them. In our thesis, when we are discussing our proposed sentiment classification model, we use the "hybrid" to refer to our method that combines topic modeling (unsupervised learning), clustering (unsupervised learning), ensemble learning and single classifier building (supervised learning) together. In the literature review section, "hybrid" is used as the general meaning. **Topic model or topic modeling**. Topic modeling refers to the task that aims at uncovering the latent topics from a collection of texts which is firstly proposed in [5]. While topic model refers to the specific model that help to achieve this goal, like Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and so on. There will be detailed explanations in Chapter 4 before we have an in-depth discussion of applying topic modeling in our approaches. **Pipeline or proposed pipeline.** In the thesis, pipeline or proposed pipeline will refer to the combination of our proposed data collection method, sentiment analysis model, and data processing details illustrated by the use case in a sequential way which covers the majority components in a typical sentiment analysis workflow that could be used as a whole. **Sentiment classification.** We have defined sentiment analysis refers to the task of extracting the sentiment orientation for the given text and assigning them to either positive or negative class, and also the techniques that could be potentially used to achieve this goal. It could contain a series of

			related tasks also before and after extracting the sentiment orientation. Therefore, we use the term sentiment classification to specifically refers to the core steps in a sentiment analysis task: data preparation, sentiment orientation extraction, and evaluation. **Comprehensive or comprehensive pipeline.** The term comprehensive usually refers to including all or nearly all components or aspects of something. In a typical workflow of Twitter sentiment analysis, at a high level, it usually contains data collection, sentiment classification, and apply the model on a use case. Since our proposed pipeline covers these three components in the thesis, we will use the word "comprehensive" to describe the trait of our pipeline that has a nearly full coverage of essential components in a Twitter sentiment analysis workflow. **Noisy labels.** In Twitter sentiment analysis studies, there are two commonly used ways to label training datasets in terms of their sentiment orientation. The first way is to label the datasets with human efforts. One or more human annotators will participant the labeling process, and if there is more than one person, their opinions on the annotations can be considered based on majority voting or based on some other schema. The other way to label the tweets is using some existing features which are called noisy labels in Twitter to categorize them into different polarities. The most commonly used one is emoji or emoticon, which is considered indicators for the sentiment of the related Twitter. Since there are no human judgements involved, and there could be noise by annotating the tweets this way so that these emojis or emoticons are called "noisy labels".

21. Why binary solution is good enough?

Before revision, we didn't explicitly mention that the sentiment analysis problem we are solving is a binary classification problem in introduction, and why we defined it as a binary classification problem.

After revision, we add a related part to provide the several reasons why we define it as a binary classification task.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	1.2 Terminology used	-	...Secondly, we want to clarify that based on our definition, sentiment analysis is a binary classification task rather than detecting the sentiment strength (one possible way is to assign a score from 1 to 5 based on the strength of the sentiment, or other score scopes, for both positive and negative). We do not consider detecting sentiment strength because of the following reasons: In order to detect the sentiment strength for the given text, training data with sentiment strength is required to contain strength as well which is hard to acquire in an automated way; An alternative could be using lexicon that has sentiment strength for each word, but lexicon-based approach performs poorly for Twitter sentiment analysis. Meanwhile, we want to compare our proposed method with previous works, and previous works mostly consider sentiment analysis as a 2-way or 3-way classification task as well. Overall, due to the goal of building an automated pipeline, and making comparisons with previous works, we want to keep our task as a binary classification problem.

37. Multiple instead of binary classification ... clarify that ...

This comment shares the same revision with comment 21. After revision, we explain why we choose to run a binary classification instead of a multi-classification in our thesis in section 1.2 Terminology used, Sentiment analysis, opinion mining, and subjectivity analysis.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	1.2 Terminology used	-	...Secondly, we want to clarify that based on our definition, sentiment analysis is a binary classification task rather than detecting the sentiment strength (one possible way is to assign a score from 1 to 5 based on the strength of the sentiment, or other score scopes, for both positive and negative). We do not consider detecting sentiment strength because of the following

			reasons: In order to detect the sentiment strength for the given text, training data with sentiment strength is required to contain strength as well which is hard to acquire in an automated way; An alternative could be using lexicon that has sentiment strength for each word, but lexicon-based approach performs poorly for Twitter sentiment analysis. Meanwhile, we want to compare our proposed method with previous works, and previous works mostly consider sentiment analysis as a 2-way or 3-way classification task as well. Overall, due to the goal of building an automated pipeline, and making comparisons with previous works, we want to keep our task as a binary classification problem.

8. Introduction also does not specify overview of existing gaps ... you should include that

Before revision, we only discussed the existing gaps in literature parts but didn't bring it up in Introduction, which may make the readers feel unclear what the gaps are when reading Introduction section.

After revision, we still have the detailed discussion of the gaps for data collection, sentiment classification, and case studies respectively, but also add an overview of the gaps for the three parts in the introduction part. Therefore, the readers can have a better idea after finishing reading Introduction. We also add a high-level summary of literature review in Introduction chapter for consistence and completeness.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
-	1.1 Background	-	Due to the availability of user-generated text online, it is also witnessed that a lot of studies have been done that specifically focus on sentiment analysis on social media, especially in Twitter by applying a variety of methods. Among these methods, we can see that lexicon based approach which is used to solve general sentiment analysis problems is introduced for Twitter sentiment analysis tasks; Supervised and semi-supervised learning based approach are two most popular methods that are used for Twitter sentiment analysis when at least a certain amount of training data is available; Also, there is a trend that different approaches are combined together for Twitter sentiment analysis in order to utilize the advantages from different methods, which makes hybrid approaches becoming more and more popular in recent years. Although it is true that a big amount of works has been done in previous works, there are still some gaps to be filled. First of all, the availability of Twitter specific training

			datasets is becoming more and more limited for the past several years, and there is a lack of researches that using graphic emojis as noisy labels when building training datasets. Secondly, for the studies that were using noisy labeled training datasets which have no topic limitations, few attentions are paid to group the tweets based on their latent semantic meanings which will hurt the classification performance due to the existence of synonymy and polysemy. Furthermore, although it is important to introduce data processing details which can help to apply the proposed model to solve real business problems in a daily basis, few previous studies are paying enough attention to this part. Therefore, we think it is essential for us to propose a comprehensive sentiment analysis pipeline which includes data collection, sentiment classification, and a use case study to provide a mostly automated, configurable system that can be used by companies or institutions in a daily basis for free. Most importantly, each component in our proposed pipeline will fill the corresponding gaps mentioned above.

12. In the introduction section, the problem statement is weak ... re-write that ...

This is a similar problem we had like the one in comment 8. Before revision, we had a weak problem statement in Introduction part which makes the readers feel unclear what's the problems we try to solve in the thesis by only reading Introduction chapter.

After revision, we add more transition between the background and the literature review by adding more details about the problems we are trying to solve and how we are going to solve them.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	1.1 Background	-	Although it is true that a big amount of works has been done in previous works, there are still some gaps to be filled. First of all, the availability of Twitter specific training datasets is becoming more and more limited for the past several years, and there is a lack of researches that using graphic emojis as noisy labels when building training datasets. Secondly, for the studies that were using noisy labeled training datasets which have no topic limitations, few attentions are paid to group the tweets based on their latent semantic meanings which will hurt the classification performance due to the existence of synonymy and polysemy. Furthermore, although it is important to introduce data processing details which can help to apply the proposed model to solve real business problems in a

			daily basis, few previous studies are paying enough attention to this part. Therefore, we think it is essential for us to propose a comprehensive sentiment analysis pipeline which includes data collection, sentiment classification, and a use case study to provide a mostly automated, configurable system that can be used by companies or institutions in a daily basis for free. Most importantly, each component in our proposed pipeline will fill the corresponding gaps mentioned above. In this study, a novel topic-model based hybrid system is introduced, which aims at solving the problems in Twitter sentiment analysis from end to end. We want to focus on the parts that have not been solved in previous studies in terms of data collection, Twitter sentiment classification, and data processing details when applying a model to solve real problems. For data collection, we propose a new automated way to collect training datasets based on graphic emojis which tries to fill the gap of currently there are few publicly available training datasets and the con of using string emoticons; For sentiment classification, we propose a topic-model based hybrid sentiment classification model which is can find the best topic model among three different topic models (LDA, HDP, and LSI) in an automated way by employing coherence score; A use case is also illustrated to explain the details about how to utilize this model for a real business problem in the perspective of data processing in a daily basis. The major goal of proposing the whole pipeline is to make most of the steps in the pipeline automated and configurable, to give it potential to be used by academics or small or medium-size companies for free.

10. Amazon, Twitter ... why mentioning the country?

Before revision, we mentioned that Twitter is a US based social network which is unnecessary in this case and not related to our task d.

After revision, we drop the “US based” phrase and rephrase the sentence.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
Introduction Page 2	1.1 Background Page 12	…Twitter, a popular social network based in the US,…	…Twitter, a popular microblogging and social networking platform,…

2. Why not mention Prophet in thesis motivation

Before revision, we didn't mention Prophet as our motivation when building the whole pipeline.

After revision, we add a motivation section that explains why we want to build such a Twitter sentiment analysis pipeline and in what scenario this pipeline should be helpful.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	1.3 Motivation	-	It is true that a lot of works have been done for Twitter sentiment analysis for the past several years. While after being inspired by Prophet , a time series forecasting package built by Facebook that aims at making forecasting at scale when solving real forecasting problems, we propose that to apply a Twitter sentiment analysis model to solve real sentiment analysis problems on Twitter, the model should be mostly automated, configurable, can be updated periodically, and include the major parts of sentiment analysis in the pipeline to keep it consistent. Making the model automated aligns with the main idea of software engineering which can reduce human effort and the probability of errors caused by human actions; Making the model configurable is also important because in the real use cases, models that easy to tune and configure are popular than the ones more like black boxes; Another thing that few previous works have discussed is how to make the model run periodically and always up-to-date, which is important when building online tools; And few works have talked about all the major components of a Twitter sentiment analysis pipeline in one study. In our study, the design of the data collection method is an essential step to make the whole pipeline automated; The design of sentiment classification model is to ensure that the components within the model are configurable; The use case aims to show how to run our pipeline online with continuously coming training data and test datasets and update the model periodically and always choose the best model for usage. By doing this, we want to make our proposed method applicable for real business problems.

18. Revise details about contributions in your thesis (what exactly is your contribution) ... Make the contributions more concrete and explicit.

Before revision, the wording for contribution in our thesis is not concise and specific enough.

After revision, we list a separate subsection for contributions under Introduction, and rewrite the bullets under contributions to make them clearer and more specific.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
1.2 Thesis Overview	1.4 Contributions	Based on all above, our main contribution could be as following: • Be the first study that utilizes graphics emojis instead of string emoticons as noisy labels for automatically annotating; • Propose a new approach to automatically determine the best topic number for across three topic models with a topic range which makes the whole process no intensive human judgment required; • Create a tweet corpus with more than 160k tweets collected via Twitter streaming API; • Build a novel hybrid system to detect the semantic topics behind the tweets by combine supervised learning, topic modeling, clustering, and lexicon knowledge together based on confidence score; • Give the system potential to be decomposed which makes it easy to tune, and to evolve by easily acquiring new data from Twitter without minor effort; • As a non-domain specific model, our proposed system achieves 78.54 for F-score; • Propose a visualization system on top of backend design which supports capturing the sentiment trend over time rather than only the trend for the past a few hours as what most of the free online tools do;	Based on the problems statements in the background section, the major goal of this thesis is to propose a comprehensive pipeline for Twitter sentiment analysis that is mostly automated and configurable, so that it can be utilized to solve real business problems in a daily basis and benefit companies and institutions as a free tool. Specifically, the following contributions are made: • Propose a new method to collect and build Twitter training datasets based on manageable and traceable graphic emojis in an automated way, without being impacted by Twitter's redistribution policy and time effect; • Propose a novel topic-model based hybrid sentiment analysis model to consume Twitter training datasets annotated by noisy labels by utilizing unsupervised learning and supervised learning approaches; • Propose an automated way to find out the best topic model without human effort when applying several different topic models with different parameters setting; • Provide data processing details when applying our proposed pipeline to solve a real business problem on a daily basis. • Propose a comprehensive pipeline including training data collection, Twitter sentiment classification, and data processing details for a real use case which is mostly automated and configurable.

29. What do you consider novel about the proposed architecture? There are a lot of other commercial tools that are doing similar things ...
It is not particularly novel ... it is not clear that how it is novel?

Before revision, it was unclear that what's the novelty part of our proposed architecture compared with previous works. We introduce the design details but have few descriptions about what the novelty is.

After revision, we add discussion sections for proposed sentiment classification model and the whole pipeline in section 4.6 Discussion and section 6.2 Pipeline analysis and discussion. The novelty in the perspective of sentiment classification model is discussed in 4.6 Discussion right after we introduce the whole model design, and the novelty in the whole pipeline's perspective is discussed in section 6.2 in a higher level. We propose the contributions mentioned in 1.4 Contributions are all novel when doing in different perspectives, but we think choose some paragraphs under section 4.6 Discussion and section 6.2 Pipeline analysis and discussion may be the best to answer this question.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
-	4.6 Discussion 6.2 Pipeline analysis and discussion	-	(4.6 Discussion) In this section, we want to discuss how our proposed topic-model based hybrid sentiment classification model is a novel model compared with the ones from previous works or from online tools. The most import feature our proposed model has is the ability to choose the best topic model from a variety of topic models with different parameters settings. For topic modeling, the number of topics is normally required for models like LDA and LSI, which can be hard to determine without a deep understanding of the dataset. In previous works, some researchers made the number of topics as a fixed number (like topic number equals to 100 for all the runs applied by the model), and others apply HDP model which doesn't require to provide the topic number before modeling. While these approaches didn't solve the problem fundamentally. For the former approach, using a fixed topic number can bring subjectivity and bias, and most importantly, there is no guarantee that the model built with the fix topic number would be the best model; For the latter approach, using HDP doesn't require the topic number while there is no way to verify the HDP model with self-determined topic numbers could outperform the LDA model with a fixed topic number. In our approach, we consider that deciding the topic number is just a way to find out the best topic model based on existing data but not the final goal, while the real goal is to find out the best topic model which can uncover the

			hidden semantic topics better than other models. Therefore, we build a pipeline in section 4.4 by employing three different topic models with different parameters setting, in order to find out the best model based on coherence scores. By doing this, there is a guarantee that the selected model for further clustering is the best one not only among the same topic model with different parameters, but also the best across the three topic models we use in the pipeline. We consider this is a novel method and one of our major contribution. In addition, the whole process is running in an automated way by using coherence score for evaluation. Although there are several topic model evaluation methods, but none of the others can help us achieve our goal: Eye Balling models require human efforts and makes it hard to compare different models in a quantitative way; Perplexity could be used in an automated pipeline but it does not consider the coherence between the examined word and the topic; To align with our goal to make the whole pipeline as automated as possible and also consider the coherence between the word and the latent topic, coherence score is a good way to run evaluation among a series of topic models being built, and also easy to be integrate with the whole pipeline. It is also possible to compare the coherence score from models with different topic numbers to have a general idea of how the coherence scores change with the different topic numbers. While for the previous works we reviewed, perplexity is commonly used as a measurement for topic models. Therefore, the use of coherence score to compare the performance among different topic models is also a novel method proposed in our study. (6.2 Pipeline analysis and discussion) … we try to make every step automated when we design data collection, sentiment classification, and the data processing steps in the use case. In previous works, few studies consider automating the whole pipeline from data collection to sentiment classification. Normally, the model they proposed is based on some human-labeled datasets which will require extra human effort if

			it needs to be updated, or some datasets built by noisy labels but with a model not designed to run periodically. As a result, the previous models seldom consider building a highly automated pipeline but require human effort for maintaining from time to time. In contrast, it is clear that the method we proposed to build training data based on Twitter Streaming API and noisy labels require nearly no human involvement. Also, the steps in sentiment classification model building are also mostly automated. For topic modeling, we utilize three different topic models and coherence scores to determine the best model in an automated way; We employ K-means clustering to assign the final cluster label to the tweets, and the only part require human judgment is to decide the number of clusters based on within-cluster sum of square, which we think is a good way to control the final number of clusters just in case a relatively big topic number is selected in the topic modeling step which would separate the training dataset into too many subgroups; For the use case, we discuss the details about how to ensure the data freshness in terms of the training dataset, classification model, and front-end dashboard and make model performance evaluation automatically. Overall, our proposed pipeline is mostly automated for all the steps which align with our final goal.

40. Clarify what is topic modeling ... it is unclear what exactly is your contribution

Before revision, topic modelling was a term that causes confusion for readers.

After revision, we firstly bring up this term in section 1.2Terminology used and briefly introduce what is topic model or topic modelling; In addition, we give detailed explanation in section 4.3 Topic Modeling Prerequisites to explain why topic model is necessary in our study and what problem it tries to solve, and also introduce 4 most popular topic models. In section 1.4 Contributions, we rewrite our contribution related to topic modeling to make it clearer and more specific.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	1.2 Terminology used 1.4 Contributions	-	(1.2 Terminology used) **Topic model or topic modeling.** Topic modeling refers to the task that aims at uncovering the latent topics

			from a collection of texts which is firstly proposed in [5]. While topic model refers to the specific model that help to achieve this goal, like Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and so on. There will be detailed explanations in Chapter 4 before we have an in-depth discussion of applying topic modeling in our approaches. (1.4 Contributions) • Propose a novel topic-model based hybrid sentiment analysis model to consume Twitter training datasets annotated by noisy labels by utilizing unsupervised learning and supervised learning approaches; • Propose an automated way to find out the best topic model without human effort when applying several different topic models with different parameters setting;

4. Some of the terms are not right terms used in your write-up. E.g., Parameterize ... it should be configurable. Also, Modular approach / component based approach is a better approach ... Flexible ... means system is not hard coded, it should be called as configurable ...

Before revision, some words are not clearly used to explain the works we did.

The revision of this comment is similar to question 3. We stop using some terms and mainly describe our pipeline as configurable and mostly automated.

41. Define all the terms clearly ... e.g., What is scalability ... define it, and clarify that how your proposed solution is scalable ...

Before revision, we used the word “scalability” to emphasis the potential of the model that could be used by different people, for different questions, with automated evaluation methods. While this is not the common meaning when people use scalability to describe a model or system and may lead to confusion.

After revision, we decide to stop using the word “scalability” or “scalable” to describe our pipeline. Firstly, it can lead to some confusions since it usually refers to the potential of a system to handle a much bigger scale of data or requests; Also, the statement of whether a model is “scalable” to different people or different questions is from the paper of Prophet, which is the study motivated us to do something similar for Twitter sentiment analysis. But we feel that after we understand their goal in that paper, we don’t have to use the same word in our thesis. Configurable, mostly automated could be the better words to describe our proposed pipeline’s feature in our study. Therefore, the sentences using

"scalability" and "scalable" to describe our system are rephrased in a more proper way now. So far, the words "scalability" or "scalable" only appear in literature related sentences.

26. What is job scheduling? provide details that what is meant by that?

Before revision, we used phrase "job scheduling" in the thesis without defining the meaning of it explicitly. What we want to describe is to run our model periodically to drive product or meet business needs in a daily basis, so that we need to schedule the jobs to trigger a new run of our model from time to time.

After revision, we stop using this phrase in Chapter 6 and rewrite the related sentences.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
2.3.1 Systems proposed by academics 2.4 Problem Statement 6.1.2.3 Data visualization Illustration 19 Diagram for Implementing Job Scheduling	…several paragraphs	…job scheduling…	*…several sentences are re-written*

34. 6.1.2.3. Visualization ... it does not say anything about visualization
Clarify that how your presented visualization is better? no mention of that in the literature review ...
Mention how there is novelty in visualization, or do not mention it as your contributions ...

Before revision, we used the term "visualization" in introduction and Chapter 6, also when claiming our contribution which is not a clear description of the work we have done.

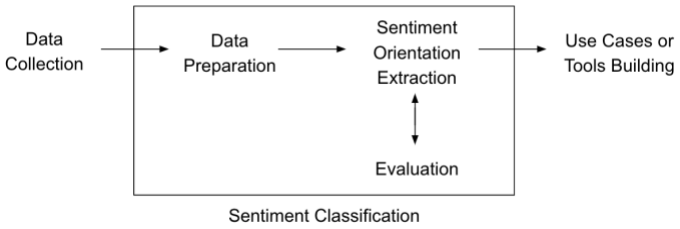
After revision, we drop the statement that "we propose a visualization system" and similar phrases in the thesis and describe our works in Chapter 6 as a use case study.

Literature Review

27. Chapter 2, you can use a chart ... or a high level workflow ...
Before revision, there was a lack of typical workflow of sentiment analysis tasks which may lead to confusion why we are reviewing the previous works for data collection, sentiment analysis and use cases and commercial tools.

After revision, we add some introduction at the beginning of the literature review chapter and also add a graph to show the typical workflow in a high level, which allows the reader to have a better idea why we review papers from these three domains.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	Chapter 2: Literature Review	-	A typical workflow or pipeline of Twitter sentiment analysis is shown in Illustration 1, which usually start with data collection, followed by data preparation, sentiment orientation extraction and evaluation, and also include the use cases or tools built on top of the previous steps. In this chapter, we will review the steps in this workflow in terms of data collections, sentiment classification, and use cases or tools building, respectively.  **Illustration 1 A Typical Workflow of Twitter Sentiment Analysis**

20. Sentiment classification is binary? Why it has to be binary? clarify that ...
For example, Emotion may be score based ...

Before revision, we are mostly talking about sentiment analysis as a 2-way or 3-way classification task, which could also be a more fine-grain task.

After revision, we add some paragraphs about the previous work that consider sentiment analysis as a task that assign a sentiment polarity and also a strength score to the given text.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	Chapter 2: Literature Review	-	Instead of regarding sentiment classification as a 2-way or 3-way classification, some research also considers it as a more fine-grain task which can potentially provide more insight to the stakeholders. In [41], the authors try to response the increasing interest in the affective dimension of the social web especially in Twitter. To fill the gap that most sentiment analysis algorithms are not ideal to this task because they utilize indirect sentiment indicators, they want to test an improved version of the SentiStrength for sentiment strength detection based on direct sentiment indicators. After testing with six social media datasets including Twitter, they found that SentiStrength not always outperform machine learning based approach, especially when it comes to detect the positive post in news; While SentiStrength does outperform baseline accuracy for positive class for all datasets and also the baseline for negative class expect for 2 datasets, which shows some extent of robustness across different datasets with different language type and style. One major contribution is that the authors applied a fine-grain sentiment analysis to response the increasing interest of it. Compared with 2-way or 3-way classification tasks which categorizes the given text into positive or negative class, this task focus on assigning sentiment strength score from 1 (weakest) to 5 (strongest) for both positive and negative classes. They also try to verify the robustness by applying SentiStrength on social network related datasets from 6 sources. It would be better if the authors can test their SentiStrength based approach on social media datasets that built on different times, since the changing language style and the newly created hashtags, buzzwords, and so on is one of the most special features compared with other text style. Therefore, it would be good to know if the performance of their proposed solution will still be robust over time.

45. Can you please suggest application areas and industries for the proposed methods? Why are the proposed methods more suitable for these applications and industries in comparison to the other existing methods? Provide more details in the in chapter 6

Before revision, there was few explanations about the suggested application areas and industries for the proposed methods, and why our proposed pipeline is more suitable for these applications or industries.

After revision, we added two paragraphs to answer these two questions respectively.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	6.1 A Use Case Study 6.2 Pipeline analysis and discussion	-	(6.1 A Use Case Study) Overall, we use a use case to illustrate how to utilize our proposed pipeline to apply Twitter sentiment analysis with "Shopify" as the keyword. It is clear that how the final results can benefit marketers in different companies and institutions by providing different keywords. For marketing purpose or maintaining public relation, marketers have a need to monitor what customers said on social media about their brand or companies. Some basic metrics including the number of views, clicks, comments, likes, shares could provide a general view about the popularity of a specific topic related the brand or company. While after applying sentiment analysis, marketers can go beyond the metrics focus on quantity, but get a deeper understanding about the loves and hates of their customers, which is a better reference before taking any further actions. Specifically, if the marketers have the opportunities to know whether their customers have positive or negative sentiment towards the product they launched, the topic they created, or the activity they held, they can have a better understanding of the preference of the customers and make adjustments timely to improve their products or services. (6.2 Pipeline analysis and discussion) Overall, it is clear that there are several benefits of using our pipeline instead of the existing tools. First of all, you can have a deep understanding of the pipeline design. For most of the existing tools, there may be documents about what methods (like, lexicon based or machine learning based) is applied in this model and how it works at a high level. While as a user, there is no space for you

			to understand the sentiment classification model driving the tools fully, nor can you explain why you get some specific results all the time. In contrast, instead of encapsulating the design details and providing the whole pipeline as a black box, our proposed pipeline is relatively understandable for every component in the pipeline. If the users are curious about why they got some specific results from the pipeline, it is possible for them to track what happened step by step within the pipeline which would facilitate a better understanding and usage of our pipeline. Furthermore, you have full control on collecting training datasets, building sentiment classification models, and applying the model in a daily basis for your problem since the whole pipeline is configurable. Some current tools don't provide any ways to make improvements when you observe some misclassification issue which means these kinds of models cannot evolve over time. While our proposed pipeline is highly configurable, which allows it to be tuned easily, and also mostly automated, which allows rapid iteration. Last but not least, the design of our pipeline considers tracking the sentiment trend, which is a missing part in almost all the free online tools. By keep tracking the tweets based on the provided keyword, our pipeline can provide a better view of the sentiment changes over time, which facilitates hourly or daily sentiment comparison and anomaly detection.

15. Clarify three main issues in gap analysis and problem statement ... re-write gap analysis and problem statement

Before revision, the gap analysis and problem statement were not clear enough to explain the gaps we are facing and the problem we are trying to solve.

After revision, we rewrite the gap analysis and problem statement for existing datasets, sentiment analysis approaches, and use case studies and tools to make them clearer and more concise. Currently the gap analysis is broken down for the three parts and sit under Chapter 2.1, 2.2 and 2.3 respectively. In this way we want to make it consistent that what papers are reviewed for the related part and what are the gaps we found from each part after literature review. While the problem statement is located in Chapter 2.4, which summarizes the overall problems we are facing when building this pipeline by claiming ideal situations, current situations, and how we plan to achieve the goal in a high-level.

The detailed revision can be found:

Position Before	**Position After**	**Paragraph/text Before Revision**	**Paragraph/text After revision**

Revision	Revision		
Chapter 2 2.1.3 Gap analysis 2.2.5 Gap analysis 2.3.3 Gap analysis 2.4 Problem Statement	Chapter 2 2.1.3 Gap analysis for existing Twitter datasets 2.2.5 Gap analysis for sentiment analysis approaches 2.3.2 Gap analysis for sentiment analysis use case studies and online tools 2.4 Problem Statement	… since the length of the paragraphs we won't paste all the text here	… since the length of the paragraphs we won't paste all the text here. Basically, we rewrite the sections listed on the left and restructure the related sections.

43. There is emphasis throughout the thesis to the point that this thesis is the first work which develops an entire ecosystem. Given the extensive literature in this area in recent years, I wonder why other works have not considered developing similar ecosystems? Provide more details/clarification about that in your thesis

This is a good question. After consideration we decide stop claiming that "this thesis is the first work which develops an entire ecosystem" but turn to describe the individual contribution for data collection, sentiment classification and case study respectively. We will explain why we claim "this thesis is the first work which develops an entire ecosystem" before and why we want to change it now:

Previously, I claimed "this thesis is the first work which develops an entire ecosystem" because of the way I collected the literatures. Sentiment analysis is a general task which mainly focuses on the sentiment classification but can also include the works before and after the classification. In order to get a better understanding of the current status for all major tasks in sentiment analysis, especially for data collection, sentiment classification, and use case studies, we didn't search for the papers that talking about the entire pipeline design but look for papers for each of the three components in the workflow specifically. One reason of doing this is we were worried that sentiment analysis is a task that including a couple of subtasks, and papers related to the entire pipeline won't go into details for each component due to the page limitation from conference and journal. In order to have a detailed and in-depth understanding of the current status for each component, we reviewed the papers for each component individually, as the structure shown now in Chapter 2: Existing Dataset, Twitter Sentiment Analysis Approaches, and use cases and tools have their own section in 2.1, 2.2, and 2.3.

Currently, we realized that due to the way we collected the literatures, the statement "this thesis is the first work which develops an entire ecosystem" is inaccurate since we didn't really put a lot of attentions on the papers that proposing the whole pipeline. There is no guarantee that no previous works have ever proposed a pipeline for Twitter sentiment analysis even without detailed explanations. Therefore, we feel that instead of saying out studies is the first one the proposing the entire pipeline, we should focus more on the contributions for the three components separately. Proposing the entire pipeline is our goal, but we shouldn't claim it as a contribution for now. Due to these considerations, we stop claiming that "this thesis is the first work which develops an entire ecosystem" but switch to claiming we made for data collection, Twitter sentiment classification, and use case study.

Proposed Data Acquisition Method

13. Size of dataset, and provide reasons and justifications that you why need a bigger and updated/recent/new dataset

Before revision, we didn't explicitly explain why it is necessary to use our proposed data collection method to building training datasets.

After revision, we add a section 3.4 Discussion under Chapter 3: Proposed Data Acquisition Method to clarify why it is essential to have our proposed method in our pipeline to server our goal.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	3.4 Discussion	-	…First of all, we are motivated to build and use this method to always collect most updated data which is due to the changing language style among tweets. In social media like Twitter, one of the most different features compared with other platforms like e-commerce or news websites it the changing language style over time. For the datasets about customer review or news, the language style is more formal and more static over time than the style in Twitter so that the time effect is not as concerned as it is in Twitter sentiment analysis. While in Twitter, new buzz words, hashtags, slangs are being created every day and evolve rapidly over time, which means the Twitter specific dataset that was valid when being used to build a model 5 years ago could no longer be valid today, since the training dataset doesn't contain the most recent tweets with the most updated language styles in the test dataset and leads to a poor performance. Therefore, we propose that in a real use case, the pipeline will be utilized to analyze the most updated tweets, so that the training dataset needs to be updated periodically as well. That is the reason why we always need more updated training dataset over time.

22. Why Emojis are so important in sentiment analysis? clarify that

Before revision, we didn't emphasis why emojis are so important in our proposed data collection method.

After revision, we add a section 3.4 Discussion under Chapter 3: Proposed Data Acquisition Method to clarify that the usage of Twitter Streaming API and graphic emojis is an essential step to make the pipeline automated.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	3.4 Discussion	-	In addition, the usage of Twitter Streaming API and emojis are essential in our method design to make the pipeline as automated as possible. When we are talking about automated sentiment analysis pipeline, we refer to the pipeline needs to have the ability to take in the newest data when it is available, utilize the data to re-build the classification model, and evaluate and apply the best model on the test dataset. Therefore, a data collection method which can keep receiving and automatically labeling the acquired data as positive or negative class is the key to drive the whole pipeline automated. So far, it is clear that why we choose to utilize Twitter Streaming API, and why we build our own graphic emoji list. The former one is to meet the need of collecting most updated tweets; The latter one is used to auto-categorize the collected tweets into positive or negative class. By doing this, we want to minimize the human effort in this process and ensure our data collection is run in an automated way.

28. Why not using graphical and text one (emojis), provide some details about that

Before revision, we mentioned that we are using graphic emojis instead of string emoticons in our study, but we didn't have an in-depth comparison of the benefits of using graphic emojis and the cons of using string emoticons.

After revision, we add a section 3.4 Discussion under Chapter 3: Proposed Data Acquisition Method to clarify that why we want to use graphic emojis and what are the benefits of using it, and also why we don't consider use string emojis in this case.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision

-	3.4 Discussion	-	Last but not least, we propose that building our own graphic emoji list is better than using string emoticons which is proposed in previous works. Our first consideration is that graphic emojis are much easier to be traced and searched, which makes it a better candidate for noisy labels. Because of the popularity of emojis, nowadays each of them has its own Unicode across different platform, which means you can search for one just like searching for any character. A variety of online sources can be found about the categorization of emojis [1] [2], and some webpages are also built to monitor the live popularity for all the emojis [3]. Therefore, using graphic emojis means that you can have a complete scope about what are all the emojis that are available on Twitter (also for any other social platform), and what is the overall volume, increasing speed, or popularity of a specific emojis. By knowing these, you can decide the collection of emojis for your use case or estimate whether the volume of selected emojis is big enough for building your dataset. In contract, using string emoticons has none of the above-mentioned benefits. Neither can you know how many string emoticons are being used by people at this moment (you can make estimations, but it is hard to exhaust all the probabilities) because there is no standard Unicode for them, nor can you estimate the volume and increasing speed of these emojis. Due to all the benefits of using graphic emojis rather than string emoticons, and there is no previous work that explicitly used graphic emojis as noisy labels, we decided to make our own graphic emoji list and apply it as a filter when collecting our training data.

1 https://emojipedia.org/unicode-10.0/
2 https://unicode.org/emoji/charts/full-emoji-list.html
3 http://emojitracker.com/

Proposed Solution for Sentiment Classification

39. Claiming that your solution is hybrid, so how exactly it is hybrid ... clarify and explain that in your thesis

Before revision, we didn't explicitly explain why our proposed sentiment classification method is a hybrid approach and why topic modelling is essential in this method.

After revision, we add some explanations to respond the two questions after we describe the overall architectural design, so that the readers can have a better understanding why our proposed model is a topic-model based hybrid approach for Twitter sentiment classification.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	4.1 Overall Architectural Design	-	About the overall architecture, we want to clarify why topic modeling is necessary and why our approach is named a hybrid approach. The reason why we want to integrate topic modeling into our pipeline is determined by the way we collect and auto-label our training data based on emojis which is introduced in Chapter 2. Because we are acquiring the training data based on graphic emojis, there will be no domain limitation on the collected tweets. If we use these tweets to build a sentiment classification model, it may suffer from poor performance due to the existence of synonymy and polysemy in the training corpus. In order to minimize the error in classification caused by a lack of domain limitation, we apply topic modeling as the first step after data preparation to capture the latent semantic topics behind the training dataset. For the reason that we call our proposed method a hybrid approach, it is because we combine unsupervised learning approaches (topic modeling, K-means clustering), supervised learning approaches (Random Forest and Logistic Regression), and lexicon knowledge (the usage of Lexicon features) together in our method. As a convention from the previous works, and as it is defined in Chapter 1.2, we name our proposed method a hybrid sentiment analysis approach. More details about each process will be discussed in the following chapters.

30. Page 102 of the thesis document (PDF file) ... clarify in subsection (Topic modeling pipeline) that how is this proposing a new system/model

Before revision, we only described how the proposed topic modelling pipeline works, but we didn't bring enough details what exactly the novel parts are.

After revision, we add a paragraph under 4.4.1 Topic modeling pipeline to clarify why this is a new model and what problem it solves compared with previous works.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	4.6 Discussion	-	In this section, we want to discuss how our proposed topic-model based hybrid sentiment classification model is a novel model compared with the ones from previous works or from online tools. The most import feature our proposed model has is the ability to choose the best topic model from a variety of topic models with different parameters settings. For topic modeling, the number of topics is normally required for models like LDA and LSI, which can be hard to determine without a deep understanding of the dataset. In previous works, some researchers made the number of topics as a fixed number (like topic number equals to 100 for all the runs applied by the model), and others apply HDP model which doesn't require to provide the topic number before modeling. While these approaches didn't solve the problem fundamentally. For the former approach, using a fixed topic number can bring subjectivity and bias, and most importantly, there is no guarantee that the model built with the fix topic number would be the best model; For the latter approach, using HDP doesn't require the topic number while there is no way to verify the HDP model with self-determined topic numbers could outperform the LDA model with a fixed topic number. In our approach, we consider that deciding the topic number is just a way to find out the best topic model based on existing data but not the final goal, while the real goal is to find out the best topic model which can uncover the hidden semantic topics better than other models. Therefore, we build a pipeline in section 4.4 by employing three different topic models with different parameters setting, in order to find out

			the best model based on coherence scores. By doing this, there is a guarantee that the selected model for further clustering is the best one not only among the same topic model with different parameters, but also the best across the three topic models we use in the pipeline. We consider this is a novel method and one of our major contribution. In addition, the whole process is running in an automated way by using coherence score for evaluation. Although there are several topic model evaluation methods, but none of the others can help us achieve our goal: Eye Balling models require human efforts and makes it hard to compare different models in a quantitative way; Perplexity could be used in an automated pipeline but it does not consider the coherence between the examined word and the topic; To align with our goal to make the whole pipeline as automated as possible and also consider the coherence between the word and the latent topic, coherence score is a good way to run evaluation among a series of topic models being built, and also easy to be integrate with the whole pipeline. It is also possible to compare the coherence score from models with different topic numbers to have a general idea of how the coherence scores change with the different topic numbers. While for the previous works we reviewed, perplexity is commonly used as a measurement for topic models. Therefore, the use of coherence score to compare the performance among different topic models is also a novel method proposed in our study.

31. Typo LMI/LSI in high level architecture ... fix that

Before revision, there was a typo in Illustration 2 Architecture of the sentiment classification system.

After revision, the LMI is corrected to LSI in the graph.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
Illustration 2 Architecture of the sentiment classification system	Illustration 2 Architecture of the sentiment classification system	… graph is too long to be pasted here…	… graph is too long to be pasted here but it is revised in the thesis…

42. Feature selection ... how do you do that? which techniques you are using? and why? Provide more details about that

After reviewing the thesis, we found that there is a section describing why and how we run feature selection on our training dataset which is in Chapter 4.2.4 Feature selection. We add more explanations to the paragraph for clarification.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
4.2.4 Feature selection	4.2.4 Feature selection	We want to run a feature selection for the N-gram features specifically since the vector space is pretty big for N-gram features and any only part of them are helpful for the model building. Based on our data exploration, we could see that a variety of infrequently used words could be seen in the dataset (88.3% of the words are being used less than 10 times while only 1.87% of the words is used more than 100 times across all the tweets) which might introduce some noise into the analysis. In order to reduce the impact of some rarely used words, we want to run a feature selection step to filter out the top n most representative ones to build our model later. Chi-square [92] method is selected for our feature selection. The way to calculate chi square is: $$X^2 = \sum \frac{(O-E)^2}{E}$$ where O refers to the observed frequency, while E refers to the expected frequency if no relationship existed between the variables.	We want to run a feature selection for the N-gram features specifically since the vector space is pretty big for N-gram features and any only part of them are helpful for the model building. Based on our data exploration, we could see that a variety of infrequently used words could be seen in the dataset (88.3% of the words are being used less than 10 times while only 1.87% of the words is used more than 100 times across all the tweets) which might introduce some noise into the analysis. In order to reduce the impact of some rarely used words, we want to run a feature selection step to filter out the top n most representative ones to build our model later. Chi-square [92] method is selected for our feature selection. The way to calculate chi square is: $$X^2 = \sum \frac{(O-E)^2}{E}$$ where O refers to the observed frequency, while E refers to the expected frequency if no relationship existed between the variables. After applying Chi-square test, only the top n (a parameter that provided by user) most statistically significant features will

			be kept in the matrix, which can help to reduce the running time of the later topic modelling, and also drop the commonly used words in both positive and negative classes, which can bring some noise later.

Sentiment Classification Evaluation

24. Describe what is F-score, cite it and provide proper reference, and describe why you need it

Before revision, we didn't define clearly the performance evaluation metrics we use in our study.

After revision, we add the formulas and explanations for all the four metrics we report in the evaluation part and explain why F-measure is a good measurement for comparison.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	5.2.1 Performance report on the proposed solution	-	For performance evaluation, accuracy, precision, recall, and F-measure are reported. We choose these four metrics since they are the most popular ones for Twitter sentiment analysis tasks which can facilitate comparison among different studies. The formulas for the four metrics are: Accuracy = $\frac{\sum True\ Positive+\sum True\ Negative}{\sum True\ Positive+\sum True\ Negative+\sum False\ Positive+\sum False\ Negative}$ Precision = $\frac{\sum True\ Positive}{\sum Predicted\ Positive}$ Recall = $\frac{\sum True\ Positive}{\sum Actual\ Positive}$ F-measure = 2 * $\frac{Precision * Recall}{Precision + Recall}$ where true Positive refers to the correctly predicted positive values when the predicted class is positive, and the actual class is also positive; True Positive refers to the correctly predicted negative values when the predicted class is negative, and the actual class is also negative; False Positive refers to the wrongly predicted negative values when the predicted class is positive but the actual class is negative; False Negative refers to the wrongly predicted positive values when the predicted class is negative but the actual class is positive; Based on the formula, accuracy refers to the portion of correctly predicted observation out of overall observations; Precision refers to the portion of correctly predicted positive observation out of overall predicted positive observations; Recall refers to the portion of correctly predicted positive observation out of overall actual positive observations; F-measure [108] is calculated by taking weighted average of Precision and Recall which is a balance between these two metrics. Compared with the

			other three metrics, F-measure is a more comprehensive metric when comparing among different models. Meanwhile, it is also commonly used for performance comparison among different sentiment analysis models when applying on the same test dataset.

32. Comparison ... same datasets were not used for comparison of results ...
Do you have a comparison of your work with existing datasets and others work with your dataset (so datasets are the same) ...
Any complications or challenges in using your dataset for learning and using your dataset for testing ... what is the difference of features ...

Before revision, we didn't explain clearly what the datasets we used as training and test datasets in the Evaluation section which may lead to confusion to readers.

After revision, we provide more details about the datasets we used: self-collected dataset as training dataset and SemEval-2013 task 2-B dataset as test dataset. We also add clarification about the limitation in comparison so far and specify the way how we compare our results with a previous work.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	5.1.1 Performance report on the proposed solution 5.2.2 Performance comparison with other works	-	(5.1.1 Performance report on the proposed solution) The training dataset we used is collected by our own based on the method discussed in Chapter 3, with 290817 tweets with positive graphic emojis and 153324 tweets with negative graphic emojis. The full positive and negative emojis we used as filter can be found in **Error! Reference source not found.**. We apply our proposed method and baseline model on Twitter in SemEval2013 Task 2-B as test dataset. There are three types of labels (positive, negative, and objective-OR-neutral) in the dataset. We only take the ones with positive or negative labels as our test dataset with 3120 positive instances and 3120 negative instances. (5.2.2 Performance comparison with other works) After comparing the usage of different features and different supervised learning classifiers, we also want to compare our proposed approach with previous works. Since we are doing a binary classification task in our study based on SemEval-

			2013 task 2-B, which provide 3 classes (positive, negative, neutral) in training and test dataset, there are not many works available for comparison. While [28] reports the results of its model both on 3-way (positive, negative or neutral) and binary classification (positive or negative) based on SemEval-2013 task 2-B as test dataset, so that it becomes a potential work we can compare our result with. It would be better to compare the classification performance if it is possible to utilize the same training dataset and test dataset, but only different sentiment classification models on top of them. While since study [28] didn't describe all the details of their proposed system (there are only 4 pages without the references including the introduction, Experimental procedure, Results, and Conclusion), it is hard for use to re-implement their system and run a side by side comparison. Therefore, our comparison is applied based on different training datasets, different features representations, different model design, but the same test dataset. The best F-measure (78.54) of our model is achieved by using N-gram features, lexicon features, and negation features driven by Random Forest algorithm which outperforms their binary classification average F-measure (77.65) for unconstrained conditions.

35. Comparison of results clarify that training datasets for existing approaches were different than for the proposed solution, but same test dataset was used

This comment shared a same revision as comment 32. Before revision, we didn't clarify that we compare our model with a previous work that using the same dataset.

After revision, we provide more details for the comparison between our model and the previous works which also uses SemEval-2013 task 2-B as their test dataset and also report the binary classification result.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	5.2.2 Performance comparison	-	After comparing the usage of different features and different supervised learning classifiers, we also want to compare our proposed approach with previous works. Since we are doing a binary

	with other works		classification task in our study based on SemEval-2013 task 2-B, which provide 3 classes (positive, negative, neutral) in training and test dataset, there are not many works available for comparison. While [28] reports the results of its model both on 3-way (positive, negative or neutral) and binary classification (positive or negative) based on SemEval-2013 task 2-B as test dataset, so that it becomes a potential work we can compare our result with. It would be better to compare the classification performance if it is possible to utilize the same training dataset and test dataset, but only different sentiment classification models on top of them. While since study [28] didn't describe all the details of their proposed system (there are only 4 pages without the references including the introduction, Experimental procedure, Results, and Conclusion), it is hard for use to re-implement their system and run a side by side comparison. Therefore, our comparison is applied based on different training datasets, different features representations, different model design, but the same test dataset. The best F-measure (78.54) of our model is achieved by using N-gram features, lexicon features, and negation features driven by Random Forest algorithm which outperforms their binary classification average F-measure (77.65) for unconstrained conditions.

A Use Case and Pipeline Discussion

25. Term like "easy to implement" ... provide details that what is meant by that?

Before revision, we use phrase “easy to implement” but what our focus on in the thesis is not really about implementation. Therefore, the wording previously was not accurate.

After revision, we drop all the phrases in the thesis that saying our proposed pipeline is “easy to implement”.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
1.2 Thesis Overview 2.1.3 Gap analysis 2.4 Problem Statement 4.2.2 Data resampling 5.2.2 Performance comparison with other works	…several paragraphs in the thesis	easy to implement	*… (rephrase or deleted from the related sentence)*

33. Chapter 6: implementation ... no implementation details ... Summarize your system (pipeline) ... add implementation details ...
Too many small subsections ... avoid that if possible. Restructure this chapter ... (pipeline review)
Section 6.1 revise it completely ...

Before revision, the title didn’t imply the actual content that discussed in this chapter, and there are a lot of subsections that repeat the content that discussed previously.

After revision, we restructure this chapter completely and add more content to it as well.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
Chapter 6	Chapter 6	…the whole chapter…	*…we make a lot of changes to this chapter, so we won’t copy and paste the content here since space is limited. Overall, the major changes include:* • *We change the title to “Chapter 6: A Use Case Study and Pipeline Discussion”;* • *We restructure all the small subsections and keep the content all in two subsections;*

			• *We add more contents in 6.2 Pipeline analysis and discussion to clarify how our proposed pipeline meet our initial goal;*

44. Do the proposed novel data acquisition and sentiment classification methods and the implementation and visualization techniques may have any commercial value? Do you consider commercialization? Provide more details in the in chapter 6

As a response to the comment, we add a paragraph at the end of Chapter 6 to discuss the consideration about commercialization.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
-	Chapter 6 6.2 Analysis and discussion	-	So far, our proposed pipeline meets all the needs and goals we planned. Previously, we have seen some models proposed by academics which have potential to achieve good performance but are hard to be applied in real problems due to its complexity to understand and tune; Also, we have seen some free commercial tools perform poorly and can hardly be used directly in business applications. We hope our proposed pipeline could benefit small and medium companies and institutions and could be used as a free pipeline to solve their real business problems.

3. Decomposable and flexible and comprehensive ... how do you define these terms and what exactly do you mean by that? Clarify that in your thesis

Before revision, we claimed our proposed pipeline is “decomposable”, “flexible”, and “comprehensive” without explicitly defining what we meant by claiming that. Also, some wording was careless which didn’t describe the exact traits our proposed pipeline has.

After revision, we decide to stop using the words “decomposable” and “flexible”. Instead, we want to emphasis our proposed pipeline is configurable and mostly automated. To support out statements, we provide details in 6.2 Pipeline analysis and discussion to explain why our proposed pipeline is configurable and mostly automated. For word “comprehensive”, we specify the meaning of it in 1.2 Terminology used.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/ text Before Revision	Paragraph/text After revision
-	Among different parts of the whole thesis; 6.2 Pipeline analysis and discussion 1.2 Terminology used.	-	(6.2 Pipeline analysis and discussion) So far, we have introduced the details of our proposed pipeline, and utilize a use case to show how to apply our pipeline to solve a business problem. In this section, we want to review the whole pipeline and discuss some main benefit of using our proposed pipeline. In the motivation, we mention that our goal is to build a sentiment analysis pipeline that is mostly automated, configurable, and can be updated periodically, which are all achieved in the pipeline design. Firstly, we try to make every step automated when we design data collection, sentiment classification, and the data processing steps in the use case. In previous works, few studies consider automating the whole pipeline from data collection to sentiment classification. Normally, the model they proposed is based on some human-labeled datasets which will require extra human effort if it needs to be updated, or some datasets built by noisy labels but with a model not designed to run periodically. As a result, the previous models seldom consider building a highly automated pipeline but require human effort for maintaining from time to time. In contrast, it is clear that the method we proposed to build training data based on Twitter Streaming API and noisy labels require nearly no human involvement. Also, the steps in sentiment classification model building are also mostly automated. For topic modeling, we utilize three different topic models and coherence scores to determine the best model in an automated way; We employ K-means clustering to assign the final cluster label to the tweets, and the only part require human judgment is to decide the number of clusters based on within-cluster sum of square, which we think is a good way to control the final number of clusters just in case a relatively big topic number is selected in the topic modeling step which would separate the training dataset into too many subgroups; For the use case, we discuss the details about how to ensure the data freshness in terms of the training dataset, classification model, and front-end dashboard and make model

			performance evaluation automatically. Overall, our proposed pipeline is mostly automated for all the steps which align with our final goal.
			Secondly, our propose pipeline is configurable, which allows it to be applied to different questions. For data collection in our study, the auto-labeling is based on 42 commonly used graphic emojis which we consider is a relatively complete list for the emojis with distinguishable sentiment orientations. While the other users can easily change this emojis collection based on their own problem, by adding more emojis or narrow down the scope. This change can be easily applied by adding or removing Unicode of the emojis that used as the filter of Twitter Streaming API as shown in Illustration 24. For sentiment classification, three topic models are used in our pipeline and for the two which require topic number as preliminary, we parameterize the minimum and maximum topic number required so that the user can adjust the topic number scope based on their own needs; For all the topic models, parameters (no_below and no_above) are provided to define the words will be included in topic modeling; As for clustering, the minimum and maximum cluster number can be decided by the use and a graph will be shown about the within-cluster sum of square of K-means for different numbers of clusters, which gives users the flexibility to determine the final number of clusters; The users can also switch between resampling, feature extraction, feature representative, feature selection modes, feature used, and classifier used. A code
			`twitter_stream.filter(`
			`track=[u"\U0001F600",`
			`u"\U0001F601",`
			`u"\U0001F603",`
			`u"\U0001F604",`
			`u"\U0001F606",`
			`u"\U0001F609",`
			`u"\U0001F60A",`
			`u"\U0001F60C",`
			`u"\U0001F60B",`
			`u"\U0001F60D",`
			`u"\U0001F60E",`
			`u"\U0001F60F",`
			`u"\U0001F617",`
			`u"\U0001F618",`
			`u"\U0001F619",`
			`u"\U0001F61A",`
			`u"\U0001F62C",`
			`u"\U0001F638",`
			`u"\U0001F63A",`
			`u"\U0001F63B",`
			`u"\U0001F63C",`
			`u"\U0001F63D"`
			`],`
			`languages=['en']`
			`)`

Illustration 2 Example Code of Applying Emojis Filter for Data Collection

```
def main(self,
         no_below=5, no_above=0.4,
         lda_min_topic_num=3, lda_max_topic_num=30,
         lsi_min_topic_num=3, lsi_max_topic_num=30,
         min_cluster_number=2, max_cluster_number=15,
         resampling_mode='r_under_s',
         feature_extraction_mode = 'unigram', bigram_min_count=10,
         feature_represent_mode='tfidf',feature_selection_mode='chi2',
         classifier = 'logistic_regression',
         show_sample_tweets_head=15,
         feature_mode ='ngram_and_lexicon'):
```

Illustration 3 Example Code of Main Function for Sentiment Classification

snippet is shown in Illustration 3. Apart from the available parameters, addition or deletion or revision to the current pipeline is also doable. For example, LDA, LSI, and HDP are used as the topic models in our pipeline. While if another topic model is expected to be added into the pipeline, it can be added without impacting any upstream and downstream. This is also applicable for changes on clustering and classification algorithms. For the front-end dashboard built based on Tableau, dashboard management and revision can be achieved without coding; For end users, filters are provided on the dashboard to focus on the parts they are most. Overall, the way we design each step is modular and loosely coupled, which allows the whole pipeline configurable.

(1.2 Terminology used) **Comprehensive or comprehensive pipeline.** The term comprehensive usually refers to including all or nearly all components or aspects of something. In a typical workflow of Twitter sentiment analysis, at a high level, it usually contains data collection, sentiment classification, and apply the model on a use case. Since our proposed pipeline covers these three components in the thesis, we will use the word “comprehensive” to describe the trait of our pipeline that has a nearly full coverage of essential components in a Twitter sentiment analysis workflow.

5. How do you show that how your system is modular and configurable? Provide details that how your system has these capabilities. Either you show in use-cases to prove it,

or you show the exact specification in your software to prove it ... not verifiable ... add more details about that

We want to respond this question similar to question 3. During the revision, we add a new section 6.2 Pipeline analysis and discussion to bring more details and explanations that why our proposed pipeline is mostly automated, configurable, and can be updated periodically in a daily basis.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
-	Among different parts of the whole thesis; 6.2 Pipeline analysis and discussion	-	So far, we have introduced the details of our proposed pipeline, and utilize a use case to show how to apply our pipeline to solve a business problem. In this section, we want to review the whole pipeline and discuss some main benefit of using our proposed pipeline. In the motivation, we mention that our goal is to build a sentiment analysis pipeline that is mostly automated, configurable, and can be updated periodically, which are all achieved in the pipeline design. Firstly, we try to make every step automated when we design data collection, sentiment classification, and the data processing steps in the use case. In previous works, few studies consider automating the whole pipeline from data collection to sentiment classification. Normally, the model they proposed is based on some human-labeled datasets which will require extra human effort if it needs to be updated, or some datasets built by noisy labels but with a model not designed to run periodically. As a result, the previous models seldom consider building a highly automated pipeline but require human effort for maintaining from time to time. In contrast, it is clear that the method we proposed to build training data based on Twitter Streaming API and noisy labels require nearly no human involvement. Also, the steps in sentiment classification model building are also mostly automated. For topic modeling, we utilize three different topic models and coherence scores to determine the best model in an automated way; We employ K-means clustering to assign the final cluster label to the tweets, and the only part require human judgment is to decide the number of clusters based on within-cluster sum of square, which we think is a good

			way to control the final number of clusters just in case a relatively big topic number is selected in the topic modeling step which would separate the training dataset into too many subgroups; For the use case, we discuss the details about how to ensure the data freshness in terms of the training dataset, classification model, and front-end dashboard and make model performance evaluation automatically. Overall, our proposed pipeline is mostly automated for all the steps which align with our final goal. Secondly, our propose pipeline is configurable, which allows it to be applied to different questions. For data collection in our study, the auto-labeling is based on 42 commonly used graphic emojis which we consider is a relatively complete list for the emojis with distinguishable sentiment orientations. While the other users can easily change this emojis collection based on their own problem, by adding more emojis or narrow down the scope. This change can be easily applied by adding or removing Unicode of the emojis that used as the filter of Twitter Streaming API as shown in Illustration 24. For sentiment classification, three topic models are used in our pipeline and for the two which require topic number as preliminary, we parameterize the minimum and maximum topic number required so that the user can adjust the topic number scope based on their own needs; For all the topic models, parameters (no_below and no_above) are provided to define the words will be included in topic modeling; As for clustering, the minimum and maximum cluster number can be decided by the use and a graph will be shown about the within-cluster sum of square of K-means for different numbers of clusters, which gives users the flexibility to determine the final number of clusters; The users can also switch between resampling, feature extraction, feature representative, feature selection modes, feature used, and classifier used. A code

```
twitter_stream.filter(
                      track=[u"\U0001F600",
                             u"\U0001F601",
                             u"\U0001F603",
                             u"\U0001F604",
                             u"\U0001F606",
                             u"\U0001F609",
                             u"\U0001F60A",
                             u"\U0001F60C",
                             u"\U0001F60B",
                             u"\U0001F60D",
                             u"\U0001F60E",
                             u"\U0001F60F",
                             u"\U0001F617",
                             u"\U0001F618",
                             u"\U0001F619",
                             u"\U0001F61A",
                             u"\U0001F62C",
                             u"\U0001F638",
                             u"\U0001F63A",
                             u"\U0001F63B",
                             u"\U0001F63C",
                             u"\U0001F63D"
                             ],
                      languages=['en']
)
```

Illustration 4 Example Code of Applying Emojis Filter for Data Collection

```
def main(self,
         no_below=5, no_above=0.4,
         lda_min_topic_num=3, lda_max_topic_num=30,
         lsi_min_topic_num=3, lsi_max_topic_num=30,
         min_cluster_number=2, max_cluster_number=15,
         resampling_mode='r_under_s',
         feature_extraction_mode = 'unigram', bigram_min_count=10,
         feature_represent_mode='tfidf',feature_selection_mode='chi2',
         classifier = 'logistic_regression',
         show_sample_tweets_head=15,
         feature_mode ='ngram_and_lexicon'):
```

Illustration 5 Example Code of Main Function for Sentiment Classification

snippet is shown in Illustration 3. Apart from the available parameters, addition or deletion or revision to the current pipeline is also doable. For example, LDA, LSI, and HDP are used as the topic models in our pipeline. While if another topic model is expected to be added into the pipeline, it can be added without impacting any upstream and downstream. This is also applicable for changes on clustering and classification algorithms. For the front-end dashboard built based on Tableau, dashboard management and revision can be achieved without coding; For end users, filters are provided on the dashboard to focus on the parts they are most. Overall, the way we design each step is modular and loosely coupled, which allows the whole pipeline configurable.

Conclusion and Future Work

36. You cannot handle any neutral tweets, so that is fundamental ... heavily relying on labels ... how do you handle neutral tweets ... you should address it in future work

Before revision, we just mentioned briefly the consideration of neutral tweets in the future work.

After revision, we clarify the current limitation and the work we plan to do in the future in terms of neutral tweets problem and providing more implementation details.

The detailed revision can be found:

Position Before Revision	**Position After Revision**	**Paragraph/text Before Revision**	**Paragraph/text After revision**
Chapter 7.2 Future work	Chapter 7.2 Future work	Another potential improvement is to include the neutral tweets into our scope and all the positive, negative, and neutral tweets into consideration. So far, the high-quality neutral data is still hard to be acquired without human efforts compared positive and negative ones which in some cases tend to be the company with emojis to express the strong sentiments. Some works try to use Wikipedia or news websites as resources to build the neutral dataset, and some try to utilize the official accounts of media. Either way of them has pros and cons.	Apart from proposing a Twitter sentiment analysis pipeline, this study also raises several questions that can be investigated further in further works. Currently, we consider sentiment analysis as a binary classification based on the pipeline, which does not consider neutral as a class in the result. While in reality, the existence of neutral tweets is not a surprise to us. We did not include neutral tweet is due to the consideration of making the whole pipeline mostly automated but acquiring the newest neutral tweets in an automated way without human efforts is still a challenging task for us at this moment. Wikipedia has been considered as a source of building neutral sentiment analysis dataset, while since the language styles between Wikipedia and Twitter have a fundamental difference, the utility of using Wikipedia need to be tested; News official accounts in Twitter is considered as another potential source for neutral tweets. The benefit of this approach is all the neutral part is built on Twitter which aligns with the positive and negative datasets; While one concern is whether the volume of neutral tweets collected this way will match the volumes

			for positive and negative tweets; And also, how to verify the selection of news accounts is good enough for building the datasets. We need to spend some time in future work to determine the best way to build neutral dataset in an automated way. In addition, we utilize a use case to illustrate how to apply our proposed pipeline on a real business problem, while the final goal is to make the whole pipeline a product which requires the user less work to run, schedule, and maintain. Currently, we only discuss the steps in the use case from the perspective of data. In further work, we will also discuss the implementation details for our proposed pipeline.

Others

38. Technical writing ... thesis ... all claims must be clearly listed in the proposed solution section, and all claims must be clearly backed-up ...

We put this comment under Others section since it related to revisions of several chapters.

Previously, some word we used to describe our proposed model or pipeline are not clearly defined or clarified. The readers may be wondering why our proposed model or pipeline has these features and how they work.

After, we define/clarify/discuss the core statements we make for our proposed methods. To be more specific, the sections we newly add and the goal of these sections are shown in the table below:

Related Part	Added Section	Goal
Proposed Data Acquisition Method	3.4 Discussion	*To clarify why having the most updated datasets are import for Twitter sentiment analysis; Why emojis play an important role in building training datasets in automated way; Why using graphic emojis is significantly better than using string emoticons;*
Proposed Solution for Sentiment Classification	4.6 Discussion	*To clarify what's the novel part of our proposed sentiment classification model compared with previous works. We discuss why the ability to choose the best topic model from a variety of topic models and the automated evaluation based on coherence scores are novel in our study.*
A Use Case and Pipeline Discussion	6.2 Pipeline analysis and discussion	*To clarify how our proposed pipeline is mostly automated, configurable, and can be updated periodically to meet real needs.*

16. A lot of subsections. Avoid too many subsections where possible ...

Before revision, there used to be a lot of subsections with five levels in the titles which is quite trivial and make the structure really deep.

After revision, we revise the related parts. Currently, the sections having four levels are the ones having the deepest levels. This could be observed directly from the table of contents or outline.

The detailed revision can be found:

Position Before Revision	Position After Revision	Paragraph/text Before Revision	Paragraph/text After revision
Several sections	Several sections	Chapter 2 previously: ▼ Chapter 2: Literature Review ▼ 2.1 Dataset for Sentiment Analysis 2.1.1 Existing datasets review 2.1.2 Literature summary 2.1.3 Gap analysis ▼ 2.2 Sentiment Analysis Approaches 2.2.1 Lexicon-based approaches ▼ 2.2.2 Machine learning based approaches ▼ 2.2.2.1 Supervised learning based approaches 2.2.2.1.1 Single classifier based approaches 2.2.2.1.2 Ensemble learning based approaches 2.2.2.1.3 Deep learning based approaches ▼ 2.2.2.2 Semi-supervised learning based approaches 2.2.2.2.1 Topic-model based approaches 2.2.2.2.2 Self-learning & co-learning 2.2.2.2.3 Graph based approaches 2.2.2.3 Unsupervised learning based approaches 2.2.3 Hybrid approaches 2.2.4 Literature summary 2.2.5 Gap analysis Chapter 6 previously: ▼ Chapter 6: Proposed Implementation and ▼ 6.1 System Description ▼ 6.1.1 Sentiment classification model building 6.1.1.1 Data collection ▼ 6.1.1.2 Model building 6.1.1.2.1 Data completeness check 6.1.1.2.2 Model building 6.1.1.2.3 Model evaluation and selection ▼ 6.1.2 Streaming data monitoring 6.1.2.1 Data collection 6.1.2.2 Data analysis 6.1.2.3 Data visualization 6.2 A Use Case	Chapter 2 currently: ▼ Chapter 2: Literature Review ▼ 2.1 Existing Dataset for Twitter Sentiment Analysis 2.1.1 Existing Twitter Datasets review 2.1.2 Literature summary 2.1.3 Gap analysis for existing Twitter datasets ▼ 2.2 Twitter Sentiment Analysis Approaches 2.2.1 Lexicon-based approaches ▼ 2.2.2 Machine learning based approaches 2.2.2.1 Supervised learning based approaches 2.2.2.2 Semi-supervised learning based approaches 2.2.2.3 Unsupervised learning based approaches 2.2.3 Hybrid approaches 2.2.4 Literature summary 2.2.5 Gap analysis for sentiment analysis approaches ▼ 2.3 Twitter Sentiment analysis use case studies and online tools 2.3.1 Use case studies and online tools review 2.3.2 Gap analysis for sentiment analysis use case studies and online tools 2.4 Problem Statement Chapter 6 currently: ▼ Chapter 6: A Use Case and Pipeline Discussion 6.1 A Use Case Study 6.2 Pipeline analysis and discussion

19. While defining existing terms and cite them properly and list the sources in references

During the revision, we review the whole thesis and check our citation or reference again when referring some existing terms, expression, or illustrations. Some of the citation or reference are not proper or missing previously, so we change and rewrite the related part, by either cite the related paper, or add footnote in the thesis.

1. Fix all Grammatical and English language errors in your thesis

After making the revisions based on the above comments, we read through the paper again and also use some grammar checking online tools to examine the potential errors in the thesis. We try our best to revise the grammar mistakes we can find.