Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science and Media

Identifying Controversial Topics in Large-scale Social Media Data

Master's Thesis

Olaoluwa Phillip Anifowose
Born Nov. 27, 1984 in Nigeria

Matriculation Number 114328

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Sven Bertel

1. Advisor: Dr. Henning Wachsmuth
2. Advisor: Michael Völske

Submission date: March 21, 2016

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, March 21, 2016

.....
Olaoluwa Phillip Anifowose

Abstract

The use of the World Wide Web has hugely impacted the way of life of men. Social media has been a platform were people from different cultures, races and ideologies interact, discuss and share different views and opinions on different issues and topics. These discussions sometimes lead to controversies between the people involved, because topics in some areas like politics, religion, history, philosophy, parenting, sex: in which people have different inclinations and opinions are well known to be controversial [Kar93] [SCRJ04]. These controversial topics are either already existing topics like topics form history which have caused controversies over the years, or it could from a topic that came up as a result of a recent event. which might lead to productive debate and discussion among the people involved, but could also lead to tension and ill feeling among the people involved. Therefore, there is the need to effectively and efficiently detect controversial topics, firstly to give people information on these controversial topics, thereby allowing them to share their views and opinions about the issue and secondly to notify the necessary authorities involved about the possible effects these topics might cause.

In this thesis, we develop a system that automatically detects Controversial topics in pages using data crawled from Reddit: a social navigation site. The data contains submissions from 2006 to 2015 and comments from 2007 to 2015. Altogether there are about 196 million submissions and 1.7 billion comments with 370 million distinct authors. We represent a page as (s, c, t) where s is the submission, c representing all comments made on the submission and t, the time the submission was created. Using this page representation, we formulate the task as a Supervised Machine Learning problem, and develop a model that classifies a page as controversial or not controversial using features adapted from existing controversy measures and some other new measures we develop. We also propose two simple methods to retrieve topics from all the pages classified as controversial. Furthermore, we also evaluate each of the measures used and see how effective they are in the classification.

After classifying the dataset using our model, the model was able to detect pages that were not originally marked as controversial and a large percentage of the pages were correctly classified . This result shows the effectiveness of the approach used in this thesis in identifying controversial topics.

Contents

1 Introduction **1**

2 Background and Related Work **6**

2.1 The Task of Controversy Detection 6

2.2 Controversy Detection Task on Different Domains 9

2.3 Reddit - A Large-scale Resource for Controversy Detection . . . 15

2.4 Summary 16

3 Dataset and Extraction Methods **17**

3.1 Justification for the use of Reddit 17

3.2 Characteristics of Reddit Dataset 18

3.3 Creating a Balanced Dataset for Use in Experiment 20

3.4 Summary 24

4 Controversy Detection: Our Approach **28**

4.1 Definitions 28

4.2 Preprocessing 30

4.3 Feature Engineering 33

4.4 Classification and Controversial Topics Detection 38

4.5 Summary 40

5 Experiment Evaluation and Results **41**

5.1 Evaluation Metrics 42

5.2 Experiment Results 43

5.3 Summary 52

6 Conclusion **56**

Bibliography **62**

Chapter 1

Introduction

The World Wide Web, an open source information space, has over the years been the primary tool for billions of people all over the world to access and share information, interact and connect to each other. This has influenced and shaped our interaction with each other and has created a fast and efficient way of accessing information and knowing what is happening across the globe within an instant. But despite the numerous advantages of the World Wide Web and the way it has improved our lives individually and collectively, finding and retrieving information efficiently and effectively from the Web has been one of the major challenges the vast amount of people involved and the large amount of information readily available has presented to us.

One of the solutions to these challenges is the use of social navigation. Social navigation refers to situations in which a user's navigation through an information space is guided and structured by the activities of others within that space. These involves using the wisdom of the crowd to find the most appropriate content being searched for. Reddit, a social navigation site, founded in 2005, and has over the past years grown to become one of the most visited online communities on the web, was primarily designed to be a front page for all websites, where web visitors can be guided to places on the web worth visiting, thereby standing as a gateway to contents available on the web. Reddit allows users, also known as Redditors, to post either links to external websites or textual contents. These submissions are grouped into several communities or areas of interest called Subreddits, and Redditors can then contribute to these submissions by writing comments on them or by up-voting or down-voting them.

On Reddit, and on any other social media platform that involves interaction and discussion between two or more people, disputes or controversies among the contributors are almost inevitable as there always seem to be clash of ideas, interests or expressed positions on some issues. This is because just like in the real life, it is difficult for everyone to agree on a fact or value. Therefore, on platforms like Reddit, which serves as a recommender to web users and on other social media platform in general, it is very necessary to effectively detect these controversial topics. One of the reason why this is necessary is because of the impact the social media has on the society. The rise of Social Media over the years has created a medium through which people's reaction to situations, policies, events and any other happenings across the globe can be known. Therefore identifying controversies or disputes on these platforms can help in shaping government policies and also help in preventing conflicts in the society before situations unravel to violence. Also, these identified controversial topics can be used to inform web user seeking information about the topic on the web on whether the topic is controversial or not, and present opposing opinions about the topic: thereby allowing them to have a balanced view of the topic. Furthermore, it could help to direct people to these topics and allow them to make their own contributions to the topic.

However, manually detecting controversial topics is a very daunting one, and it is not trivially identifiable, because of the following reasons:

- Controversy is difficult to define as it could mean dispute, bias, truth value among many other things.
- The scope or context of controversy could play a major role in determining the controversy of a content. For example what is regarded as controversial differs for different people.
- The large amount of content makes it infeasible to manually check all content on the web to determine if they are controversial or not
- The diverse content on the web also makes this task difficult, as the user can not have enough expertise on all subjects.
- The dynamic nature of the contents also make the task very challenging, because the rate at which contents evolve over time is high and it might be difficult to keep track of the all these changes.

Therefore it is imperative to identify these controversial topics automatically. To achieve this, several methods are adopted by several platforms. For example, Reddit explores the voting feature of their platform by creating an

algorithm that aggregates the up-votes and the down-votes of submissions and comments and use it to determine if a page is controversial or not. But because the up-votes and the down-votes were visible to the users, it allowed users to manipulate the votes to favour the controversy score of the submissions and comments. To minimize this, Reddit has a system, which constantly identifies these accounts and cancels the votes made by them and balance the votes out and make them to represent the reality[1]. Recently, Reddit made some changes to this, [2] by making the up-votes and the down-votes no longer visible to users, but obviously using the up-votes and down-votes seems not to be the best way to detect and classify submissions as they can be easily manipulated by the users. This shows the need for an automated way of detecting controversies that is not dependent on humans.

To achieve this, there have been different research focused into detecting Controversies in pages in different domains on the web. These approaches can be categorized into the following groups.

Statistical Method Approach

This involves the use of statistical models based on the analysis of the domain used. A method in this approach is to use the metadata of the domain to determine if the page is controversial or not. For example because of the richness of Wikipedia metadata, statistics like: the number of authors, the number of page revisions were used by [VLS+08] to detect Controversial Articles in Wikipedia.

Another method is applying statistics and probability on the extracted content of the page. For example finding the percentage of verbs in the page and using the result to determine if a page is controversial or not.

Rule Based Approach

The rule based approach develops algorithms to mathematically determine how controversial a page is. For this approach two methods can be adopted. The first method is the use of the domain's metadata to compute algorithms. An example of this is the use of the Mutual Reinforcement Principle by [VLS+08] to develop a model to rank how controversial a Wikipedia page is.

The second method involves the use of the page's content in computing the algorithms that is used to know how controversial a page is. [BKSV15] for

[1] http://marketersguidetoreddit.com/why-you-should-never-buy-upvotes-reddit/
[2] Reddit changes: individual up/down vote counts no longer visible

Table 1.1: Table showing different approaches with examples of where they were used.

	Metadata Features	Content Features
Statistical Method Approach	Number of revisions on a Wikipedia Page [VLS+08].	Percentage of Verbs in twitter page [PP10].
Rule Based Approach	Controversial Rank Model [VLS+08].	Controversy Detection Algorithm by [BKSV15]

example used the content and context of the edits in a Wikipedia page to detect if the page is controversial or not controversial and how controversial they are.

Building upon this approaches, in this thesis, we develop a supervised machine learning model, using the different controversy detection measures used above, created new measures and use these measures as feature to classify our model. We use Reddit as our dataset because it contains discussion from a wide variety and diversity of people and topics which represents what happens in real life, and also because it has a rich metadata.

To make our task adequately represent and simulate what happens on Reddit and how Reddit works, we represent a page as: $p = (s, c, t)$, where s represents the submission for the page, these submissions can either be a link or text content, c represents all the comments associated with the submission and t represents the UTC time the submission was created. Representing a page like this helps us to capture all contents that are related to each other and also in classifying them.

In this thesis, our objective is to answer the following research questions:

1. How difficult is the task of finding controversial topic in a large domain.
2. How fast can these controversial topics be detected.
3. How effective are the known controversial measures in detecting controversial topics.

We believe that this work will be of help in detecting controversial topics and can also be used by search engines when presenting the results of the informa-

tion need of users. In addition, this work will also help in future research on controversies detection.

Thesis Organization

The remainder of the thesis is organized as follows: Chapter 2 reviews related scientific work and the background for our work. There we discuss extensively research regarding controversy detection, describe various approaches and point out the major differences of these approaches to the one presented in this thesis. Chapter 3 focuses on the dataset: the justification for the use of it, its characteristics, how a subset of the dataset is extracted for our experiment. In Chapter 4 we present the approach we use in this work, while Chapter 5 describes our experiment and presents our results. Chapter 6 summarizes our work and concludes with an outlook to future research.

Chapter 2

Background and Related Work

In this chapter we introduce the necessary background to the remaining part of this thesis, we first look at the task of controversy detection, giving a working definition of what controversy is and looking at some works related to our work after which, we look at various research done in detecting controversies that were done on specific domains. For all these works, we describe the various algorithms and methods presented which closely correspond to the ideas presented in this thesis, discuss their results and point out the major differences to the approach in this work. Finally, we give a brief description of the dataset that would be used in this thesis and discuss some research that has been done using it.

2.1 The Task of Controversy Detection

As discussed in Chapter 1, the task of automatically detecting controversies is a necessary one. This is because manually detecting them is not trivial. This has led to some significant number of research that have been done in this area. Controversy is defined as follows: *a state of prolonged public dispute or debate, usually concerning a matter of conflicting opinion or point of view*[1]. In these section we will take a look at some of the research done in detecting controversy in general and also discuss some of the work done that are not domain specific, but can be used across different domains.

[1]https://en.wikipedia.org/wiki/Controversy

Dori-Hacohen et al [DHYTA15] in their work looked at the challenges involved in getting information about controversial topics using search engines. They explained that because of the effect information has on the choices people make, the search task and the process of presenting the "correct" information to users becomes complex when the user's information need relates to controversial topics. They pointed out that firstly, the search engine must recognize that the query submitted by the user is controversial and also secondly, that the search engine must also determine what is controversial about it before presenting it to the user. Doing this might be very challenging and some of the challenges in developing a search engine that does this effectively and efficiently as pointed out by them include:

- The ambiguity in defining what is controversial or not controversial, as it is challenging to achieve a common agreement on this.
- The dataset or domains used might lead to defining controversies in a problematic way. An example of this is representing vandalism as controversy, and rating podcast as the most controversial topic on Wikipedia.
- Does the scope and context of the controversy play a major role in determining if a topic is controversial.
- And more importantly, disagreement among research on whether sentiment metric: a measure that aims to identify and extract subjective information in source materials, should be relevant for controversy or not, as opinions on movies or products could contain sentiments, but lack controversy.

This shows that detecting controversial topic is a complex and non-trivial task, as well as the need to find an efficient and effective way of developing an automatic way of detecting controversial topics. It also shows the need for evaluating the different available measures and finding out how well they perform in detecting controversies

However, Tsytsarau et al [TPD] in their work focused on the problem of finding sentiment-based contradictions at a large scale. While sentiment analysis can be used to reveal or determine people's reaction to an event or a topic, contradiction or sentiment diversity can be used to reveal the diverse and conflicting opinions and views people have about the event or topic. They developed a measure that aggregates the mean and the variance of sentiment, with the

sentiments extracted at sentence-level. In addition to this, they also looked at the performance and effectiveness of their measure on a large scale dataset.

From the result of their experiments of evaluating the performance of their measure on synthetic dataset, real world dataset and as well as through user-study, all the contradictions detected by their measure corresponds to discussions expressing different points of view and opinions on the same topic, The result also shows almost the same performance with or without neutral sentiments. The measure used in this work was adapted as part of the feature for our machine learning model.

Shiri Dori-Hacohen and James Allan on the other hand in their work [DHA13], and later in [DHA15] proposed a weakly-supervised approach for detecting controversy on the web, using the nearest neighbour classifier[2] to map arbitrary web pages to Wikipedia articles related to them. Given a web page, they tried to determine if the page is controversial or non-controversial. The following steps were done to solve this problem:

- They extracted the top ten most frequent words from the page excluding stop words. The words are then used as the keyword query for a search engine restricted only to the Wikipedia domain. The Wikipedia articles returned are considered as the web page's neighbours.
- The Wikipedia articles found as neigbours are then labeled with several scores that measures their controversy levels. The scores are:
  - **D Score:** This score represent the presence of dispute tags provided by Wikipedia, added to the page by contributors.
  - **C Score:** A score that uses a variety of Wikipedia metadata features (e.g length of the page, number of editors etc.) to predict the controversy level using regression as described by Kittur [KSPC07].
  - **M Score:** This score estimates the controversy level of the Wikipedia articles based on the concept of mutual reverts and edit wars [YSR+12].
- The score of the web page is then computed by either taking the maximum score or the average score of its Wikipedia neighbours.
- The score of the page could also be computed by voting using one of the voting schemes they presented.

[2] https://en.wikipedia.org/wiki/Nearest$_n eighbour_c lassifiers$

Applying this on their dataset containing 377 webpages and 8,755 Wikipedia articles, they claim that their approach achieves absolute gains of 22% in $F_{0.5}$ [HTF08], over a sentiment-based approach. Although this approach is domain independent, but the reliance on Wikipedia pages in computing the score, might make the result dependent on the quality of the Wikipedia pages.

2.2 Controversy Detection Task on Different Domains

To further look into the task of controversy detection, we take a survey of other approaches used in other research that were done using some specific domains, analyze them and discuss their results.

Controversy Detection on Wikipedia Articles

Many of the research done in finding controversies has focused on controversy detection on Wikipedia. Wikipedia, a free multilingual Internet encyclopedia, written collaboratively by different contributors, contain rich metadata and revision history that provides a valuable resource for detecting controversies. To detect controversial articles, Wikipedia allows contributors to manually tag articles as controversial. This is inefficient because of the sparseness of the tagging of pages as controversial by users, and also since not all the articles will be checked to determine if they are controversial or not for the purpose of tagging by users, therefore the justification for finding an automated way of detecting controversial articles.

Voung et al [VLS+08] in their research proposed three models drawing clues derived from the articles' edit history and collaboration among users. Considering disputes between two pairs of editors as the number of words that were written by one editor and later deleted by the other, they proposed a simple model referred to in their paper as the basic model, which measures the ratio of deletes to all contributors to the article. The other two models: Controversial Rank Model and the Age Aware Model, are based on the Mutual Reinforcement Principle. The principle as used by them can be summarized in two points:

- An article is more controversial if it contains more disputes from less

controversial contributors

- A contributor is more controversial if he had been involved in more disputes in less controversial articles.

Based on this principle, at each step of computation, the controversy score of an article is updated by the amount of dispute that happened between its editors weighted by their controversy score at that step. Next, the controversy score of editors will be updated based on the updated controversy of edited articles, and this dual updating process continues until the two scores converge. Furthermore, because articles go through a lot of review and edit when they are newly created and could be mistaken for controversy, the Age Aware Model was designed to prevent these frequent alterations to be mistaken as disputes.

To find out how their models perform, the models were evaluated on a collection of more than 19,000 Wikipedia articles and they used the number of revisions the article has gone through (Revision Count) and the number of contributors to the article (Contributor Count) as a base model to compare their results. From their experiments, the Age Aware Controversial Models was able to detect six articles that were tagged by Wikipedia users as controversial in the top 20 articles returned. They also found out that the revision counts and the contributor counts are not reliable indicators of controversy in comparison to their models. One major challenge with their approach is that the it focuses only on one category. Also the computational cost Controversy Rank and Age Aware Models becomes very high with a larger dataset.

Another work by Kittur et al [KSPC07], which examined the growth of conflict and coordination in Wikipedia and described tools in characterizing them, used machine learning in identifying the level of conflict in an article, by using the page metrics such as Revisions, Minor edits, Unique editors etc. Their goal was to predict the Controversial Revision Count (CRC) scores: the count of the total number of revisions in which the *controversial* tag was applied to the article, and to see if these statistics about the history of the document were enough to identify its level of conflict. Using about 58 million Wikipedia articles, they trained their model using only pages that are controversial in the latest revision. Their result shows that a learned model was very effective at predicting Controversial Revision Count (CRC) scores from the page metrics.

However, Bykau et al [BKSV15], tried to detect not only if a Wikipedia article is controversial or not, but also give information about when and where the controversy appeared, the topic and the author involved in it. They proposed an algorithm that cluster edits throughout the revision history of a page and

identify controversies. Their algorithm can be summarized in these steps:

1. They performed edit extraction to identify the edits that have taken place in the history of the page by comparing consecutive documents in the history using the Myer's Algorithm [Mye86].
2. From the edits generated, the eliminated all edits that have not been repeated by at least two users.
3. The remaining edits that are about the same subject are grouped together using a clustering algorithm. And each group represents a controversy.
4. To ensure that each group of edits represent only one controversy, the group of edits gotten from the step above are clustered using the Jacaard similarity[3] to find edits that are common to two different groups.
5. The found edits are then ranked based on their level of controversy using the cardinality(number of users), duration(how long the controversy lasted) and plurality(how many repetition). Using Wikipedia dump[4] from 2013 to 2015, they showed that their approach has higher precision and recall than the baseline approach and also detected previously unknown controversies.

However, Sepehri et al [RB12] in their work compared five different methods used for modelling and identifying controversies on Wikipedia domain. The five methods they compared rely on features extracted from the revision history of the page or the article discussion page, but do not analyze the textual content of the page. Some of the methods described earlier in this section like the Basic Model by [VLS+08] and the method by [KSPC07] were part of the methods compared in their work and using the same dataset of 240 controversial articles and 240 non-controversial articles for all the different methods compared, they considered three metrics: Discriminative Power (the effectiveness to distinguish controversial from non controversial articles), Training Cost (the effect of the training data on the accuracy of the methods) and Monotonicity (if it assigns less o equal score to the page if some parts of the article was removed).

From the results, they concluded that in identifying controversies, several different factors should be considered and not only using a single heuristic. Furthermore, they noted that the relative performance of all the methods remained

[3]https://en.wikipedia.org/wiki/Jaccard_index
[4]https://dumps.wikimedia.org/

the same, regardless of the amount of the training data. They also found out that most of the methods did not satisfy the Monotonicity criterion.

In summary, the approaches in these section relies very much on Wikipedia metadata and it might be difficult transferring then to other domains, especially to domains with sparse metadata. Also, some of the work, for example, the work done by [BKSV15] explores the peculiarity of Wikipedia articles which involve editing of the different segment of a page among several individuals which is only peculiar to Wikipedia articles.

Controversy Detection in Social Media Data

Some few works has also been done on controversy detection using social media data like Twitter. The social media, a platform where people interact and share information presents a rich source of data to use to detect controversies. One of such works is the work done by Ana-Maria Popescu and Marco Pennacchiotti [PP10] which focused on controversies involving celebrities. They defined a snapshot as $s = (e, \Delta t, tweets)$, where e is an entity contained in a list of Celebrities, Δt is a 1-day period and $tweets$ are the sets of tweets in the time period mentioning the entity. Using a large set of features, they used regression machine learning models to detect levels of controversies of each snapshot. Furthermore, given a set of entities and a set of snapshots, they ranked the snapshots according to a controversy detection function that assigns higher score to controversial-event snapshots and lower score to non-controversial-event Some of the features used by them in their model can be used across different domains, while some are peculiar only to Twitter. Table 2.1 shows a list of their feature sets, pointing out the ones that we also use in our experiment. From their experiments, using a set of 104,713 celebrities, they reported an average precision of 66%.

This approach is related to the work done in this thesis in that a supervised machine learning model is also designed using features extracted from the dataset and some of the features used in this approach were also used in our work. However, while they represented their work as a regression problem that predicts the score of each snapshots, we represent our work as a classification problem that predicts the class of each page using knowledge from the dataset. Another difference between their work and the work done in this thesis is that while their work focused on a selected few celebrities and snapshots within a day, our work covers a larger set and looks at the the whole lifespan of the post. Moreover, our approach also gives insight into how fast controversy

Table 2.1: Feature List used by [PP10]

Family	Feaures	Not Adapted
Linguistic	Percentage of tokens that are nouns	Adapted
	Percentage of tokens that are verbs.	Adapted
	Percentage of tokens that are bad words	Adapted
	Percentage of tweets containing at least one question	Adapted
	Average Levenshtein distance between tweets.	Not Adapted
	Percentage of tokens which match any word in the English dictionary	Not Adapted
	Average number of mentions of the target entity across all tweets.	Not Adapted
	Percentage of verbs whose corresponding subject is the target entity.	Not Adapted
	Percentage of tweets containing at least one verb whose subject is the target entity.	Not Adapted
Structural	Number of tokens in the snapshot.	Adapted
	Number of tweets in the snapshot.	Adapted
	Percentage of tweets that are retweets.	Adapted
	Percentage of tweets that are replies.	Adapted
	Average number of tweets per user.	Adapted
	Two features, representing mean and std.dev. of the distribution modeling tweets' timestamps.	Adapted
	Number of unique hashtags with respect to the total number of hashtags.	Not Adapted
Sentiment	Fraction of positive tweets.	Adapted
	Fraction of negative tweets.	Adapted
	Fraction of neutral tweets.	Adapted

Buzziness	Estimates entity buziness	Not Adapted
Controversy	Estimate how many mixed positive and negative tweets are in the snapshot	Adapted
	The contradiction score.	Adapted
	Four features, representing the fraction over the total number of hashtags in the snapshot, of the following hashtags: '#controv', '#scandal', '#unheard' and '#wft'.	Not Adapted
	Percentage of tweets with least one controversy word.	Not Adapted
News buzz	Number of articles aligned with the given snapshot.	Not Adapted
	Change in the amount of news coverage for the given entity with respect to the recent past	Not Adapted
Web-News	Controversy level of an entity in Web data.	Not Adapted
	Sum of overall controversy scores for the entities co-occurring with the target entity in the aligned news article set.	Not Adapted
	Average of overall controversy scores for the entities co-occurring with the target entity in the aligned news article set.	Not Adapted
	Average number of controversy terms per news article (over all articles aligned with the snapshot)	Not Adapted
	Max number of controversy terms per news article (over all articles aligned with the snapshot).	Not Adapted
	Number of articles aligned with the snapshot that contain controversy terms.	Not Adapted

can be detected, evaluates some known controversy measures and also identify controversial topics from the list of pages predicted as controversial.

2.3 Reddit - A Large-scale Resource for Controversy Detection

Reddit is an online social media community that allows user to post content and also vote on these post or other ones. These post are grouped into different communities known as *Subreddits* which can be created by users and are moderated by volunteers. Reddit consists of rich metadata that can be downloaded and used to conduct research.

There have been some scientific research done using the Reddit as the dataset. Notable among them is the research done by Gilbert [Gil13] which looked at the effect of underprovision on social voting sites. He described underprovison as a scenario when too many people rely on others to contribute without doing so themselves. On Reddit this can happen when people just visit and get the information needed without contributing to the submission by either voting or commenting. To evaluate this, Gilbert used statistics of the page view and the analysis of duplicate submissions: He defined duplicate submissions as submissions with links which became popular but were earlier submitted by someone else. From his research, he posited that widespread underprovisioning of votes occurs on the site and that 52% of the most popular sites were overlooked the first time they were submitted. This shows that relying on users activity in determining controversies may not be reliable as there is a possibility of a post not being voted for because of low human activity, although it is controversial. This therefore shows the need to find a way of detecting controversies automatically without depending on human activity.

Singer et al [SFM+14] on the other hand in their work looked at the evolution of Reddit, they looked at how user submissions have evolved over time and how the perception and attention of the community towards submissions has changed. They analyzed 60 million submissions and from the result, they found out that Reddit has increasingly transformed to a self-referential community that focuses on its own content over external sources.

Table 2.2 shows the user study results of the percentage of participants agreeing to a description of Reddit and Table 2.3 shows the percentage of participants using features of Reddit. They also found out that there was an exponential

Table 2.2: Participants description of Reddit

How would you characterize Reddit?	Percentage
Forum / Message board	88
Entertainment site	71
News site	56
Image/Video or file sharing site	54
Portal	48
Educational site	43
Social Network	33
Other	26

Table 2.3: Participants use of Reddit features.

Users said they...	Percentage
...never/seldom submit content to Reddit.(n=669)	78
...often or very often vote on submissions. (n=670)	55
...often or very often comment on submissions. (n=665)	32

increase in the number of submissions over the time considered.

This result further confirms the decision in using Reddit as the dataset for use in this work, as it is a good source to get textual contents with diverse areas of interest and also a good representation of discussions and interactions happening in the real world.

2.4 Summary

This chapter introduced the necessary background to the work done in this thesis presented in subsequent chapters. We looked at the domain we intend to use and later did a survey on the research done in detecting controversial topics looking at different approaches and some domains. In Chapter 3 we will take a closer look at our dataset, the characteristics of the dataset and how we extract a subset for use in our experiment.

Chapter 3

Dataset and Extraction Methods

Controversy as defined in the previous chapter is a state of prolonged public dispute or debate, usually concerning a matter of conflicting opinion or point of view. Therefore, to identify these public disputes or debates, a domain that is a true representation of what happens in the real world among people interacting together is needed. This is one of the criteria used in choosing a suitable dataset for the experiments done in this thesis. In this chapter we introduce the dataset that is used in this work and discuss the justification for using it. Later on in the chapter some statistics are presented to understand the characteristics of the dataset and finally the method used for extracting the subset of the data used for the experiments are explained.

3.1 Justification for the use of Reddit

We use Reddit as the domain for the experiments in this thesis. Reddit, as earlier described, is an on-line social media platform designed to help guide users to places worth visiting on the web. Reddit allows registered members to submit content such as links or text contents known as *selftext*. These submitted contents called *submissions* are organized into different areas of interest or communities known as *Subreddits* and registered users can respond to these submissions by either expressing their opinion or reaction on these submissions or by voting for or against the submission. In Reddit, posts that expresses opinion or reaction on a submission are known as *comments*, while the voting can either be an *upvotes* or a *downvotes*

The large number of people in the Reddit community and the high level of interaction among them, makes Reddit a suitable domain for controversy detection. As of February, 2016, Reddit had more than 542 million visitors monthly and rank as the 13th most visited website in the United States and 27th in the world[1]. Throughout 2015, Reddit had 82.54 billion page views, 73.15 million submissions, 725.85 million comments made by 8.7 million total authors containing 19.36 billion words, 88,700 active Subreddits and 6.89 billion upvotes from its users[2].

Furthermore, another advantage of using Reddit is that, Reddit has an algorithm that measures how controversial a *comment* is. and includes in the metadata a value that indicates if a comment is controversial or non-controversial. This value in the metadata gives an additional information which is used later in this thesis, and also provides a means to test the performance of our approach.

3.2 Characteristics of Reddit Dataset

Reddit data consisting of all submissions posted to Reddit from July 2006 to May 2015 and all the comments from October 2007 to May 2015 crawled through the Reddit's API [3] is used in this work. This Reddit data, which is in JSON format and crawled by[4] [5], contains 196,531,736 submissions and 1,659,361,605 comments. Table 3.1 presents some of the fields of the submissions' schema as well as the comments' schema that are later used in this work.

To better understand the composition of the dataset, some analysis are carried out to find the statistics of the dataset. Table 3.2 shows some major statistics of the dataset. In addition to this Figure 3.1 shows the cumulative frequency distribution of the number of comments whose controversial field value is one for each submissions. Considering only submissions that have at most 50 of such comments, 50.36% of these submissions have only one controversial comment, 67.22% submissions have at most two controversial comments and 90.48% of the submissions have at most eight controversial comments.

[1] http://www.similarweb.com/website/reddit.comoverview
[2] http://www.redditblog.com/2015/12/reddit-in-2015.html
[3] https://github.com/reddit/reddit/wiki/API
[4] https://www.reddit.com/r/datasets/comments/3mg812/full-reddit-submission
[5] https://www.reddit.com/r/datasets/comments/3bxlg7/full-reddit-comment

Table 3.1: Selected Fields from the Dataset Schema

Field	Description
approved_by	Representing the moderator that approved the comment
author	Representing the author of the post (submission or comment)
banned_by	Representing the moderator who banned the comment
body	Representing the content of the comment.
controversiality	Representing if the comment is controversial
created	Representing the authors time when the post was submitted
created_utc	Representing the server time when the post was submitted
disable_comments	Representing if comments are disabled for the submission
downs	Representing the down vote for the post
gilded	Representing if gold was bought for the post
id	Representing the unique identifier of the post
link_id	Representing the submission that the comment belong to.
over_18	Representing if the submission is for people above the the age of 18
parent_id	Representing the post to which the comment is responding to
removal_reason	Representing the details of the removal of the comment
selftext	Representing the content of the submission (Empty if the submission is not a text content).
Subreddit	Representing the area of interest where the submission belongs.
Subreddit_id	Representing the Subreddit id
ups	Representing the up vote for the post
url	Representing the posts' locator

Table 3.2: Important Statistics on Whole Dataset

Statistics	Number
Number of Submissions	196,531,736
Number of Comments	1,659,361,605
Number of Subreddits	430,482
Number of Submissions' author	14,644,634
Number of Comments' author	13,213,173
Average of words in submissions	10.735
Maximum of words in submissions	599
Average of words in comments	37.506
Maximum of words in comments	37,955
Number of Controversial Submissions	2767225
Number of Non-Controversial Submissions	193775971

Furthermore, taking a look at the composition of Subreddits in the dataset, Table 3.3 shows the top 20 Subreddits with the highest number of submissions and Table 3.5 shows the top 15 Subreddits with the highest number of comments, while Table 3.4 shows the top ten submissions with the highest number of comments. In the top ten Subreddits in Table 3.4 three of those submissions belong to the *AskReddit* Subreddit. This Subreddit as shown in 3.4 also has the highest number of submissions and also the highest number of comments as seen in Table 3.5, which is more than three times the number of comments in the second Subreddit with the highest number of comments.

3.3 Creating a Balanced Dataset for Use in Experiment

One of the goals of this thesis is to detect controversial pages from a large scale dataset. Using the information from the dataset, a Controversial Comment is assumed as a comment whose *controversiality* field is equal to one, while a Controversial Submission is assumed as a submission that has at least one Controversial Comment and a Non-Controversial Submission as a submission without a Controversial Comment. Furthermore as defined in previous chapter, a page is represented as $p = (s, c, t)$ where s represents the submission for the page, which can either be a link or text content, c represents all the comments

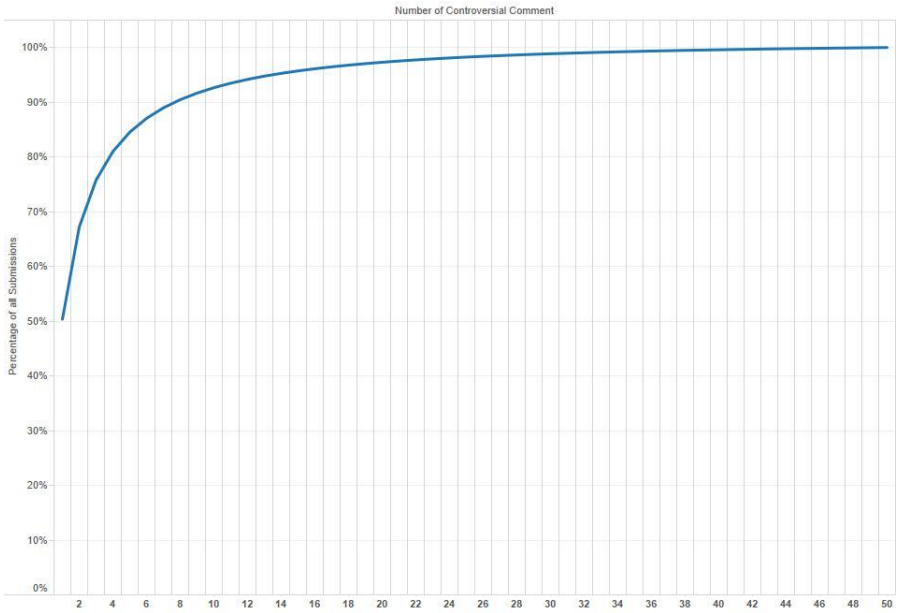

Figure 3.1: Cumulative frequency distribution of the number of controversial comments per submissions

associated with the submission and t represents the time the submission was created in UTC. Using this page representation and the assumptions above, we define a page as controversial if its submission is controversial and a page as non-controversial if its submission is non-controversial.

Based on our definition of what a page is, in this work, a model is designed that predicts if a page is controversial or non-controversial, after which experiments are done to check how the model performs based on the assumptions given above. For this experiment and other experiments like the one in this thesis, it is necessary to have balanced set of pages, i.e equal number of controversial and non-controversial pages. This set of data should also be balanced in the diffrent areas of interest and time considered. This is necessary to prevent any form of bias in the prediction. In order to achieve this balance, the *subreddit* and the *created_utc* field of the Reddit dataset is used. The choice of these two fields is because the *subreddit* represents the areas of interest and the *created_utc* represents the time, which will ensure that the dataset used for experiments is balanced across the same areas of interest and the same time period between the the controversial and non-controversial pages to be considered.

Figure 3.2 illustrates how this is done and the process of extraction is described in the following steps presented below.

Table 3.3: List of Top 20 Subreddit with the highest number of Submissions

Subreddit	Number of Submissions
AskReddit	65231
funny	60122
leagueoflegends	40353
AdviceAnimals	38911
worldnews	36520
pics	34769
gaming	33113
nba	33064
WTF	32701
soccer	27660
videos	26888
todayilearned	25908
politics	25117
DotA2	24843
nfl	20813
SquaredCircle	19554
news	19002
movies	16252
starcraft	15051
Bitcoin	14250

1. From the collection of all submissions, all the submissions that have at least one comment tagged as controversial is extracted as Controversial Submissions while the remaining set of submissions from the collection are extracted as Non-Controversial Submissions.

2. For each set of these submissions, all submissions without comments and Subreddits were removed. This step is done as a clean up because as explained above, the Subreddit is one of the categories used in ensuring a balanced dataset. Therefore a Submission without Subreddit is not useful for the experiment

3. The set of Controversial Submissions are then grouped by their Subreddit and time (Month and Year). This produces a list of pairs of Subreddit and time for all the Controversial Submissions, and in addition to this, the number of submissions that belongs to each pair.

Table 3.4: Top 10 submissions with the highest number of Comments and their Subreddit

Submission Id	Subreddit	Number of Comments
d14xg	blog	358913
28sluw	jerktalkdiamond	146676
2q36z6	millionairemakers	108959
3g4blw	AskReddit	59165
zi4f1	circlejerk	57400
2jcc98	DeadlyEighteen	56287
2syfcu	millionairemakers	51023
1witn6	AskReddit	50287
2kjze6	podemos	45536
t0ynr	AskReddit	45147

4. The List produced from the above step is then used to extract equal numbers of Non-Controversial Submissions from the Non-Controversial Submissions Collection using the pair of Subreddit and time from the list. The reason for using this list from the Controversial Submissions to extract equal numbers of Non-Controversial Submissions, and not the other way round, is because as shown in the statistics in Table 3.2 the number of non-controversial submissions is more than the number of controversial submissions

5. After extracting the Non-Controversial Submissions there are possibilities that the numbers of Non-Controversial submissions extracted for some pairs of Subreddit and time might not be equal to the numbers of Controversial Submissions for those pair. Also, it is possible that there might not be any Non-Controversial Submission for some pairs. Therefore, to balance the two sets of Submissions, Step three was also done on the Non-Controversial Submissions and the list generated was used to extract equal numbers of Controversial Submissions as done in Step four above.

6. From the two balanced sets of Non-Controversial Submissions and Controversial Submissions, all comments that are linked to these submissions are extracted into Controversial Submissions' Comment and Non-Controversial Submissions' Comment respectively.

Table 3.6 shows some important statistics of the extracted dataset. In total

Table 3.5: Top 15 Subreddit with the highest number of Comments

Subreddit	Number of Comments
AskReddit	184540520
funny	49967219
pics	47208205
gaming	32202209
WTF	29779435
leagueoflegends	29706965
AdviceAnimals	27339965
politics	22904996
videos	21235667
worldnews	19687581
todayilearned	18508173
IAmA	18298109
nfl	17116723
atheism	16138248
trees	14312315

there are 2,755,219 submissions in each of the controversial and non-controversial submissions, which are in 14,294 unique Subreddits that occur in 55 different month and year pairs. Also from Table 3.7 showing the top 20 Subreddits and Time pairs with the highest number of submissions, only three distinct Subreddits are in the top 20 and they occur between the year 2013 and 2015. In addition, Table 3.8 shows a representation of the total number of submissions in each of the Controversial and Non-Controversial Submissions.

3.4 Summary

In this chapter we looked at the Reddit dataset used in this work. The large amount of data and the number of people interacting on this platform gives us the opportunity to model what happens among humans in real life. We also performed some experiments to understand the characteristics of the data, and presented some of the results. In summary, the dataset gives us a good starting ground for our experiment and the dataset extraction gives us a balanced subset of the whole data. This extracted dataset will be referred to as the dataset in the remaining part of this thesis.

Table 3.6: Important Statistics from Extracted Dataset

Statistics	Number
Number of Controversial Submissions	2,755,219
Number of Non-Controversial Submissions	2,755,219
Number of Controversial Submissions' Comment	339,347,096
Number of Non-Controversial Submissions' Comment	30,171,191
Number of Subreddits	14,294
Number of Controversial Submissions' author	982,902
Number of Non-Controversial Submissions' author	1,076,907
Number of Controversial Submissions' Comment author	5,546,074
Number of Non-Controversial Submissions' Comment author	26,228,761

Table 3.7: Top 20 Subreddits and Time Pairs

Subreddit	Month_Year	Number of Submissions
AskReddit	3_2013	5529
funny	3_2013	5191
AskReddit	7_2014	4670
nba	7_2014	4599
AskReddit	4_2014	4584
AskReddit	3_2014	4565
AskReddit	5_2015	4522
AskReddit	5_2014	4487
AskReddit	4_2015	4461
worldnews	7_2014	4325
leagueoflegends	5_2015	4255
funny	3_2014	4253
funny	1_2014	4217
AskReddit	6_2014	4215
AskReddit	1_2014	4143
AskReddit	4_2013	4021
funny	4_2014	3941
AskReddit	2_2014	3840
funny	2_2014	3825
funny	4_2013	3813

Table 3.8: Representation of the number of submissions in extracted dataset

Subreddit	Time	Number of Non-Controversial Submissions	Number of Controversial Submissions
sub_1	m_y_1	n_{11}	n_{11}
	.	.	.
	.	.	.
	.	.	.
	m_y_j	n_{1j}	n_{1j}
.	.	.	.
.	.	.	.
.	.	.	.
sub_i	m_y_i	n_{ii}	n_{ii}
	.	.	.
	.	.	.
	.	.	.
	m_y_j	n_{ij}	n_{ij}

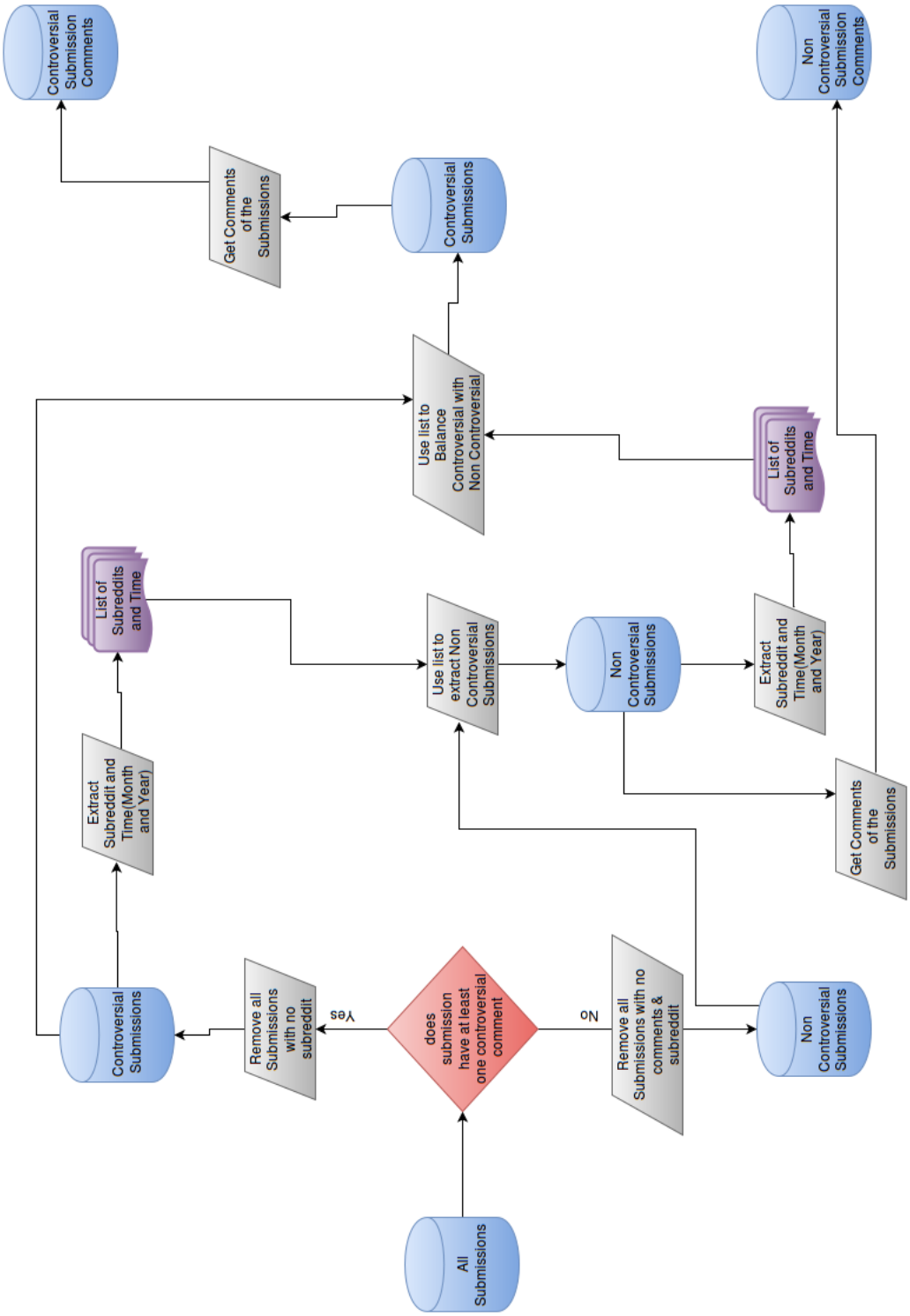

Figure 3.2: Image Illustrating the Dataset Extraction

Chapter 4

Controversy Detection: Our Approach

In this thesis, the task is to automatically identify controversial topics in a large-scale social media data, The objectives are to find how difficult this task is, evaluate some existing controversy detection measures and also find out how fast these controversial topics can be identified. To achieve this, the task is represented as a Supervised Machine Learning Problem that produces an inferred function based on the knowledge from the older part of the dataset and uses this function to make predictions on the more recent part of the dataset. In this regard, using the extracted dataset explained in Chapter 3, the dataset is organized by ordering them in an ascending manner using the page's time t, which is the time the submission was created. Using this ordering, the first 60% of the pages are used as the training set to train the model to produce this function, the next 20% are used as the validation set to test how well the function performs, and then make predictions using this function on the remaining pages that are yet to be seen, to know if they are controversial or non-controversial.

4.1 Definitions

Before going into the details of the approach used in this work, it is necessary to formally introduce some important definitions that will be needed in the rest of this thesis.

Definition 4.1.1 *Controversial Comments: A Controversial Comment is a comment in which the value of the controversiality field is one in the dataset.*

Definition 4.1.2 *Controversial Submissions: A Controversial Submission is a submission with at least one controversial comment.*

Definition 4.1.3 *Page: A page is defined as triple $p = (s, t, c)$, where s represents the page's submission, c representing all the comments that are associated with the page and t the time the page's submission was created in UTC.*

Definition 4.1.4 *Controversial Page: A Controversial Page is a page in which the submission s is controversial.*

Definition 4.1.5 *Controversial Authors: An author is controversial if he has submitted at least two comments, and the ratio of controversial comments to all the comments he has submitted is greater than 0.5.*

Definition 4.1.6 *Post: A post is a generic name representing either a comment or a submission.*

Definition 4.1.7 *Time Segment(t_s):Time segments are specific period within the dataset. The periods considered in this thesis are 0 day, 0.1 day, 0.2 day, 0.3 day, 0.4 day, 0.5 day, 1 day, 1.5 days, 2 days, 2.5 days, 3 days, 3.5 days, 4 days, 4.5 days, and 5 days, These periods are the difference between the time a submission is created and the time its comments are created.*

Definition 4.1.8 *Time Segment's Comments: The comments within a time segment are all the comments created after the last time segment to the current time segment. For a time segment t_s_i, the Time Segment's Comments at this time segment t_s_i, are all comments after t_s_{i-1} to t_s_i, where i are all the time periods in consideration.*

Definition 4.1.9 *Age Segment's Comments(a_S): The comments within an age segment are all the comments created from the first time segment to the current time segment. For a time segment t_s_i, the Age Segment's Comments are all comments from t_s_0 to t_s_i, where i are all the time periods in consideration.*

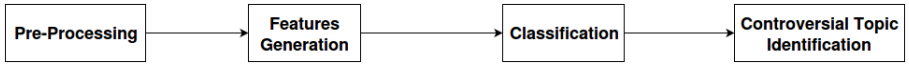

Figure 4.1: Overview of our Approach

Definition 4.1.10 *Page's Age: The age of a page is the difference between the page's time t, and the time the last comment for the page was created.*

Figure 4.1, shows an overview of the approach, the first stage involves preprocessing the dataset to retrieve all necessary elements needed in the subsequent stages. At the next stage, features are generated from the dataset: these are different characteristics of the dataset that can help in predicting a page as controversial or non-controversial. Furthermore, at the next stage, predictions are done to classify the pages as either controversial or non-controversial and this classification is evaluated. Lastly, at the last stage, from the pages predicted as controversial, controversial topics are extracted.

4.2 Preprocessing

At this stage the dataset is prepared and necessary elements of the dataset are extracted, which are used in subsequent stages. The details of what is done at each step of this stage are explained below.

Controversial Authors

At this step the list of all controversial authors is extracted from the dataset, according to the definition of Controversial Authors above.

Grouping Comments into Age Segments

From the dataset all the comments are grouped into different collections based on Age Segments as defined above. That is, for each Age Segment, the Age Segment's Comments are extracted and each Age Segment's Comments form a group

From Figure 4.2 showing the statistics of the comments per submissions at each time segments, it can be seen that the most numbers of comments per submissions are posted from 0.4 day to 0.5 day time segment and this reduces by more than half from 0.5 day to 1 day. Also it can be seen that the number

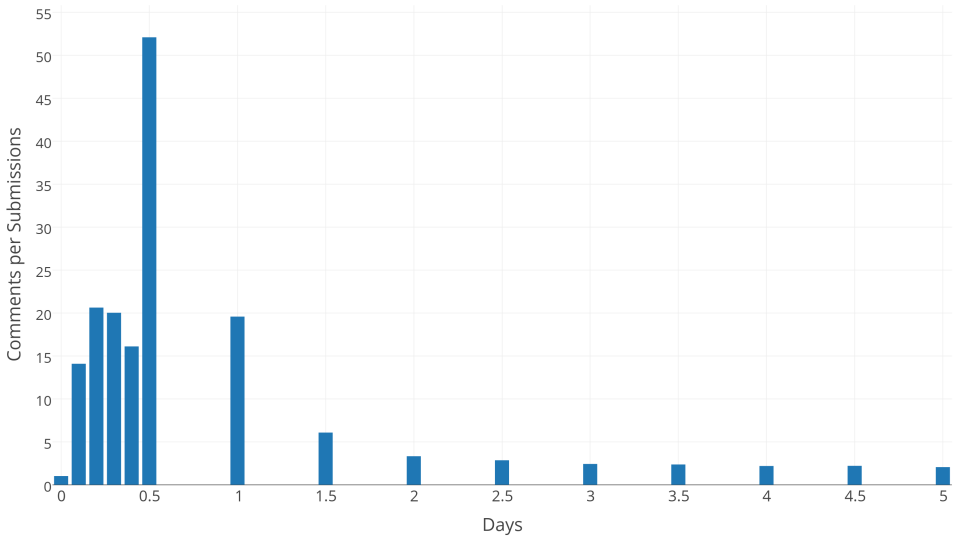

Figure 4.2: Rates of Comments per Time Segment

of comments per submissions balances out to about 2 comments per submission after the 1 day. It is also interesting to note that that there are some comments created at the time the submission was created, on further investigation into the reason for this, it was discovered that these comments are automatically generated by *automatic moderator* at the time the submission was created. This is done for example when an author who is below the age of 18 creates a submission in a Subreddit marked as *over_18*. This analysis of the number of comments per submissions in each time segment forms the reason why comments posted up till 5 days from the time the submission was created are considered in this thesis.

In addition to the comments that are grouped based on their Age Segments, there is also a collection of all comments of the page irrespective of the time difference between the time they were created and the submission's was created.

Dividing Pages based on Age Segment's into Collections

At the end of the above step, there are 16 different collections of comments which are: the first 15 comments' collections based on the Age Segments and the 16th collection that has all comments irrespective of the Age Segment they belong to. For each of these comments collection, pages are extracted according

to the definition of Pages above, that is, all the comments related to the page in each of the collections and their respective submission are extracted. This gives 16 representations of each page in the dataset.

Splitting Pages into Controversial or Non-Controversial

Using the definition of *Controversial Page* above, each of pages the pages extracted in each of the Comments' collection from the above step are then divided into Controversial or Non-Controversial Pages.

Page Ordering

At this step, each set of Controversial and Non-Controversial pages extracted from the comments' collection are ordered according to the Page's time t, in ascending order.

Tokenization and Part of Speech Tagging

After the above process, we have 16 collections of pages that contains each page's submissions and all the comments attached to them at each comments' collection. As the next process, for each of the collection of pages, the text content of the page's submission and all the comments associated with the page in that collection are broken down into words and symbols. This process is known as tokenization and each of the words and symbols are called tokens. The tokens resulting from this process is then tagged with their respective Part of Speech (POS) [1]. This is done for all the comments and only for submissions that are *selftext*. For pages whose submissions are not *selftext*, only the page's comments are tokenized and tagged with their Part of Speech.

List of Bad or Swear Words

Using Google's "What Do You Love" web service project[2]: a new search tool that combines all of Google's services into one page, and provides a portal for users to access all these services based on the their search query, a list of 450 words that are not allowed on this service[3] are used as the list of Swear or Bad Words.

List of Thank or Positive Words

For a list of positive words, the list of positive English words extracted from

[1] http://www.ling.upenn.edu/courses/Fall$_2$003/$ling$001/$penn_treebank_pos.html$

[2] http://www.wdyl.com/

[3] https://gist.github.com/jamiew/1112488

Positive Words Vocabulary List[4] are used.

Sentiment Score

Using the SentiWordNet API [BES10] [ES06], the sentiment score sent(w) of each word in the page is computed. Before computing the sentiment score, each word is first lemmatized[5] the to its root word and the sentiment score is looked up in the WordNet dictionary [Mil95] [Fel98]. The lemmatization ensures that there are more words found in the dictionary. The polarity of each post pol(p) in the page is then calculated by adding all the sentiment score for each word in the page. $pol(p) = \sum sent(w)$

4.3 Feature Engineering

A wide variety of resources is used to derive a large set of features for the model designed in this work. These features are the characteristics of the data that will help the model in predicting if a page is controversial or non-controversial. Furthermore, as stated at the beginning of this chapter, one of the objectives of this work is to evaluate some of the existing controversy detection measures and see how well they perform, because of this, some of these existing measures that can be adapted to our domain are used as features for the model. In addition to this some new features are also created which are specific to the Reddit domain.

All the features used in this thesis can be grouped into these major categories.

Structural Features

These are features that are derived from the the way the data is composed and the relationship between them. They are mainly derived from the metadata of the dataset. They include:

averageTimestamp: The average timestamp of all comments in the page.

avgNumOfContributorsInPage: The average number of distinct authors in the page using the data's authors field.

[4]http://www.enchantedlearning.com/wordlist/positivewords.shtml
[5]https://en.wikipedia.org/wiki/Lemmatisation

avgNumOfControversialContributorsInPage: Using the list of all controversial authors extracted in the preprocessing stage, this represents the average number of controversial contributors in the page.

controversialAuthorsInPagePercentage: The ratio of controversial authors to all authors in the page.

depth: The depth of the page's comments' tree.

isAuthorDeleted: A feature that returns 1 if the author of the submission has been deleted and 0 otherwise.

isSubmissionByControversialAuthor: A feature that returns 1 if the submission was submitted by a controversial author and 0 otherwise.

numOfComments: The total number of comments associated to the page.

numOfGildes: The total number of gildes in the page. A gild is gold bought for the author of a post.

percentageOfDeletedAuthors: The percentage of authors in the page that are deleted.

percentageOfGildeds: The percentage of posts in the page that has one or more gilds.

percentageOfReplies: The percentage of comments that are a reply to another comment in the page.

stdDevTimestamp: The standard deviation of the timestamp of all the comments in the page.

totalNumberOfWords: The total number of words in the page. This includes the contents of all the comments and the submission.

Linguistic Features

This set of features describes the text content of the entire page, which includes all the comments associated with the page and also the content of the submission, if the submission is *selftext*. The tokenization and the Part of of Speech tags in the preprocessing stage is used to compute these features. They include:

percentageOfAdverbs: The percentage of all the tokens in the page that are adverbs.

percentageOfBadWords: The percentage of all the tokens in the page that are in the list of bad words.

percentageOfIsInPastTense: The percentage of all the tokens in the page that are past tense.

percentageOfModals: The percentage of all the tokens in the page that are modal verbs.

percentageOfNouns: The percentage of all the tokens in the page that are nouns.

percentageOfPersonalPronouns: The percentage of all the tokens in the page that are personal pronouns.

percentageOfPunctuations: The percentage of all the tokens in the page that are punctuations.

percentageOfQuestions: The percentage of the posts (comments and subms-sion) that has a question mark in the page.

percentageOfSarcastics: On Reddit a sarcastic comment or post ends with */s* characters. This feature calculates the percentage of the posts (comments and submssion) that has this sarcastic characters in the page.

percentageOfVerbs: This calculates the percentage of all the tokens in the page that are verbs.

wordProportion: This aggregates the number of thank words and the number of bad or swear words in the page. Using the following formula:

$$\left(\frac{\sum Bad}{\sum Bad + \sum Good} + \frac{\sum Good}{\sum Bad + \sum Good}\right) \cdot \frac{1}{\sum Bad + \sum Good}$$

where Bad and Good represents the bad or swear words and thank or positive words explained above.

Sentiment Features

This set of features uses the sentiment score computed in the preprocessing stage to compute features that reveals the views and opinions expressed by the different authors of the page content. They include:

contradictionScore: The contradiction score. Adopted from [TPD]

$$\frac{\theta \cdot \sigma^2}{\theta + (\mu)^2} \cdot W$$

where μ and σ^2 are respectively the mean and the variance of the polarity scores pol(p) of the post: parameters θ and W are set as in [TPD]

controversialMix: Estimates how many mixed positive and negative posts are in the page. Adopted from [PP10]

$$\frac{\min(|Pos|, |Neg|)}{\max(|Pos|, |Neg|)} \cdot \frac{|Pos| + |Neg|}{|Pos| + |Neg| + |Neu|}$$

where Pos, Neg and Neu are the sets of posts with positive, negative and neutral polarity.

negativeFraction: The ratio of all posts whose polarity is negative to the total number of posts in the page. (i.e. $pol(p) < 0$)

neutralFraction: The ratio of all posts whose polarity is neutral to the total number of posts in the page. (i.e. $pol(p) = 0$)

positiveFraction: The ratio of all posts whose polarity is positive to the total number of posts in the page. submission. (i.e. $pol(p) > 0$)

Age Dependent Features

These time based features are dynamic and uses the age of the page in their computation. The Time Segment mentioned above is used in computing these features. The features in this category include:

commentsPerAge: This calculates the ratio of the number of comments in the page to the Page's Age

maximumNumOfCommentsInTimeSequences: Represents the number of comments in the Time Segment with the highest number of comments.

Table 4.1: Overview of the Features used in our work

Family	Features
Structural	averageTimestamp
	avgNumOfContributorsInPage
	avgNumOfControversialContributorsInPage
	controversialAuthorsInPagePercentage
	depth
	isAuthorDeleted
	isSubmissionByControversialAuthor
	numOfComments
	numOfGildes
	percentageOfDeletedAuthors
	percentageOfGildeds
	percentageOfReplies
	stdDevTimestamp
	totalNumberOfWords
Linguistic	percentageOfAdverbs
	percentageOfBadWords
	percentageOfIsInPastTense
	percentageOfModals
	percentageOfNouns
	percentageOfPersonalPronouns
	percentageOfPunctuations
	percentageOfQuestions
	percentageOfSarcastics
	percentageOfVerbs
	wordProportion
Sentiment	contradictionScore
	controversialMix
	negativeFraction
	neutralFraction
	positiveFraction
Age Dependent	commentsPerAge
	maximumNumOfCommentsInTimeSequences

Table 4.2: Time Frame of the Dataset Divisions

Set	Start Date (GMT)	End Date(GMT)
Controversial Training	18 Jun 2008 22:01:57	18 Jun 2014 16:44:37
Controversial Validation	18 Jun 2014 16:45:00	30 Mar 2015 22:26:09
Controversial Test	30 Mar 2015 22:26:11	31 May 2015 23:54:35
Non-Controversial Training	12 Jun 2008 23:12:11	18 Jun 2014 18:42:49
Non-Controversial Validation	18 Jun 2014 18:42:52	28 Mar 2015 16:22:36
Non-Controversial Test	28 Mar 2015 16:22:58	31 May 2015 23:58:02

Table 4.3: Overview of the number of pages in the dataset

Set	Total Number of Pages
Controversial Training	1,653,131
Controversial Validation	551,044
Controversial Test	551,044
Non-Controversial Training	1,653,131
Non-Controversial Validation	551,044
Non-Controversial Test	551,044

4.4 Classification and Controversial Topics Detection

As mention earlier in this chapter, the task is modeled as a Supervised Learning Problem that predicts whether a page from a set of recent data is controversial or non-controversial, using the function inferred based on the knowledge from pages from older data. To achieve this, using the ordered pages of each the controversial and non-controversial pages from the preprocessing stage, the ordered pages are divided into three sets. Table 4.2 and 4.3shows statistics of these sets.

These three sets which they are divided into are:

Training Set

The training data set make up the first 60% of the ordered pages of each of the controversial and non-controversial pages. They are the set of data which is used to build up the prediction function and the prediction function is tuned to the peculiarity of the training data sets.

Validation Set

The validation data set make up the next 20% of the ordered pages of each of the controversial and non-controversial pages. This set of data is used to compare the performances of the prediction function that is created based on the training set and used to tune the parameters of the function for better performance.

Test Set

The test data set make up the last 20% of the ordered pages of each of the controversial and non-controversial pages. After tuning the performance of the prediction function, the chosen prediction function is applied on the test set to see how the prediction function performs on an unseen real-world data.

Classification

Each page in the sets of data described above is called an instance, and using the features from the earlier section, a list of attributes are generated for each instance of the dataset. As a final step before the classification, each instance is labelled as either controversial or non-controversial, using the definition of a controversial page as described above. This labelling is to serve as the ground truth for the classification. After which the model is built with the training set and predictions are made on the validation set and the test set. The Validation Set is used to find out how the model is performing and and to know if changes are required based on the results of the classification and the Test Set is a new set of unseen data that shows how the model will perform with new sets of data. This classification is done using the WEKA tool [HFH+09] which contains several implemented classifiers.

Furthermore, due to the large number of our instances and also because of the large amount of features used in describing each of the instances in the data

set a classifier that incrementally trains the model is needed[6] Therefore for this work, the Naive Bayes Classifier is used.

Controversial Topics Detection

From the classification stage above, we get for each page, the score that shows the level of confidence that the page is correctly predicted as either controversial or non-controversial. This score which is the probability of correctly predicting the class of the page, ranges from zero to one. Therefore, to get the list of the most controversial topics, we considered only the pages whose probability of predicting the page as controversial is equal to one. Thus allowing us to only select pages with 100% level of confidence and also fewer pages.

From the list of selected pages, the title of the page is extracted using the title field of the page's submission. These titles are the list of all controversial topics detected. Also, the first ten Subreddits with the highest number of pages is also extracted and these Subreddits are the most controversial areas of interest.

4.5 Summary

In this chapter we look at our approach in detecting controversies and the extraction of controversial topics. First of all, at the beginning, we gave some useful definitions that we use. Then we went through a list of steps we took in preparation for the other stages. In the third section we look into the features, used in describing our data. these features ranges from some of the already developed measures in controversy detection that are can be easily adapted to our domain to some new measures developed in this thesis. The last section describes our approach in detecting if a page is controversial or not and also our approach in detecting controversial topics from all the pages classified as controversial.

[6]https://weka.wikispaces.com/Classifying+large+datasets

Chapter 5

Experiment Evaluation and Results

This chapter describes all the experiments done in this thesis and presents their results. Using the extracted dataset as discussed in Chapter 3, different experiments are conducted to evaluate how well the classifier can predict the pages as controversial or non-controversial, how much time from the page's time does the classifier need to be able to effectively predict these pages and also to find out the effect of time dependent training set on predicting the pages. Furthermore, the experiments also evaluates the performance of each of the features used in describing the data. The experiments are conducted on a total of 5,510,438 pages, containing equal numbers of controversial and non-controversial pages.

All the experiments carried out in this work can be categorized into three major categories. The categorization is based on the comments that are considered for each of the pages in the training, validation and test set. The experiments' category are:

Time Independent Experiments:

This experiment considers all the comments of the pages in the training and test sets regardless of the Age Segment they belong to. That is, for each page in the set, all the comments associated with the page are considered.

Time Dependent Experiments:

This experiment only considers comment within an Age Segment. Therefore,

for each page in both the training and test set, only the Age Segment's Comments for that time segment are considered as the comments for the page.

Hybrid Experiments:

The experiments combines the characteristics of the two experiments above. This experiment is done with the time independent training set and time dependent test set.

5.1 Evaluation Metrics

The evaluation metrics used in this thesis, measures the classification performance and also the the performance of the features used in describing the dataset. All the experiments above are evaluated based on the the following evaluation measures:

True Positives (TP): This denotes the number of pages predicted as controversial and were initially labelled as controversial.

False Positives (FP): This denotes the number of pages predicted as non-controversial and were initially labelled as controversial.

True Negatives (TN): This denotes the number of pages predicted as non-controversial and were initially labelled as non-controversial.

False Negatives (FN): This denotes the number of pages predicted as controversial and were initially labelled as non-controversial.

Precision: The Precision is the ratio of the True Positives to the sum of the True Positives and False Positives.

$$Precision = \frac{TP}{TP + FP}$$

It can be thought of as a measure of a classifiers exactness. The precision tries to answer this question: Out of all the pages that the classifier predicts to be relevant, how many are truly controversial? A low precision means that there are many pages predicted as non-controversial that were initially labelled as controversial.

Recall: The Recall is the ratio of the True Positives to the sum of the True Positives and the False Negatives.

$$Recall = \frac{TP}{TP + FN}$$

It can be thought of as a measure of the classifier's completeness. The recall tries to answer this question: Out of all the pages that are truly controversial, how many are can the classifier predict to be controversial? A low recall means that there are many pages predicted as controversial that were initially labelled as non-controversial.

F-Score The F-Score is a measure that combines the precision and the recall. It is the harmonic mean of the precision and the recall.

$$F - Score = 2 \cdot \frac{precision + recall}{precision \cdot recall}$$

Percentage Correctly Classified (PCC): This is the percentage of pages for which a correct prediction was made in relation to the label.

Information Gain (IG): To evaluate the performance of each of the feature, the information gain of each feature is calculated. The Information Gain of the feature t_k over the class c_i is the reduction in uncertainty about the value of c_i when the value of t_k is known. In general terms the information gain estimates the worth of an feature by measuring the information gain with respect to the class. In order words the Information Gain tells us which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.

The value of the Information Gain ranges from zero to one, with the best feature being the feature with the highest Information Gain and the worst feature, the feature with the lowest Information Gain. We use the Information Gain because it is an effective metric in measuring the performance of features used in classification as pointed out in the research by [YP97].

5.2 Experiment Results

Age Independent Experiment

The age independent experiments as explained above are experiments using all the comments of the page regardless of the age segment they belong to.

Table 5.1: Time Independent Experiment Validation Data Result

Metric	Value
True Positives	364241.0
False Positives	87057.0
True Negatives	463987.0
False Negatives	186803.0
% Correctly Predicted	75.2
Precision	0.81
Recall	0.66
F-Score	0.73

The purpose of this test is to evaluate the effectiveness of our approach in detecting controversies. In addition to this, the experiment is done to know how the features used in classification in predicting the pages using all the comments of the pages in the training set. Also from the controversial topics extracted, we want to know the effectiveness of the classifier in predicting the 10 most frequent *Subreddits* from pages from which these topics are extracted. To do this, all the metrics explained above are used. Table 5.1 and Table 5.2 presents the results of the evaluation of the Validation set and Test set. Comparing these results, Figure 5.1 shows that the classifier was able to predict a significant large number of pages as controversial which were labeled as non-controversial (True Negatives) and this number increases from the validation set to the test set. Also the number of pages predicted as controversial that were labeled as controversial (True Positives) is high: although, the number reduces from the validation set to the test set. The effect of this result can be seen in Figure 5.2, as the Precision is higher than the Recall and the Precision increases from the validation set to Test set. This means that the number of pages classified as controversial that are truly controversial are more in the test set than in the validation set which shows the effectiveness of the classifier.

Table 5.4 however presents the result of the ranking of the Information Gain on each of the features for the Test set in this experiment. The result shows that the numOfComments: a measure that counts the number of comments in a page outperforms all the other features. Also the maximumNumOfCommentsInTimeSequence, which is one of the Age Dependent Features, performs reasonably well being the second feature with the highest value of information gain. It is also instructive to note that among all the features in the Sentiment Features Family, the controversialMix feature performs best and the percent-

Table 5.2: Time Independent Experiment Test Data Result

Metric	Value
True Positives	348214.0
False Positives	68196.0
True Negatives	482848.0
False Negatives	202830.0
% Correctly Predicted	75.4
Precision	0.84
Recall	0.63
F-Score	0.72

ageOfQuestions feature among the Linguistic Features family. Lastly, seven features have scores less than 0.1, these features are: sSubmissionByControversialAuthor, commentsPerAge, percentageOfSarcastics, averageTimestamp, percentageOfGilded, isAuthorDeleted, numOfGildes. It is also important to note that none of these features belong to the Sentiment Features Family, which shows that all the features in the Sentiment Family contributed significantly to the classification.

Furthermore, from the ten most frequent Subreddit from all the pages from which the controversial topics are extracted from, Table 5.3 shows the percentage of pages correctly classified for each of these Subreedit, it can be seen that the percentage of pages correctly classified for each of these Subreddit are almost the same with the percentage of pages correctly classified for all Subreddits, apart from the the *funny* and *nfl* Subreddits which have significant higher percentage of pages correctly classified. This behaviour is expected, as about 10% of all the pages in the test set belongs to these Subreddits

Age Dependent Experiment

The time dependent experiments as explained above considers only the Age Segment's Comments at each age segment for each page of the training, validation and test set of the page. The objective of this experiment is to know how fast controversies can be detected, looking at the Precision, Recall, F-Score and the Percentage of Pages Correctly Classified. The results presented here shows the outcome of this experiment at each Age Segment of the test set.

Figure 5.3 shows the percentage of pages that are correctly predicted at each of

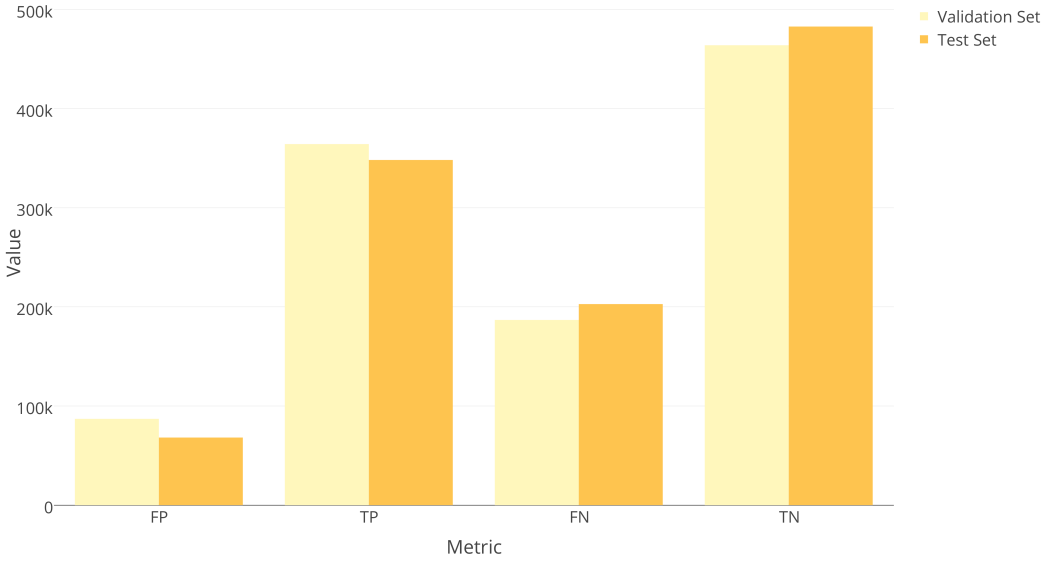

Figure 5.1: Positives Negatives Metric Result for Validation and Test set

the time frame. At 0 day, about 51% of the pages are classified correctly, this is to be expected, because at this point the only comment in this time segment are comments created automatically by the moderators and so the classifier does not have sufficient information to classify the page as controversial or not controversial. The percentage of pages correctly classified increases to 62% at 0.1 day: which is 2 hours 24 minutes after the page is created, and continues to increase until 2 days, after which the rate at which the percentage of pages correctly predicted increases, is almost zero. This time is the time the classifier has sufficient information for its classification. In real life, using the statistics from the rates of comments per submissions for each Time Segment in Figure 4.2, at this time there about two comments per submissions, showing that most comments for the page has already been submitted

Figure 5.4 shows the Precision, Recall and F-Score, The Recall at the 0 day age segment is almost one. This means that the number of pages classified as controversial that were initially labeled as non-controversial (false negatives) at this point is very low compared to the number of pages classified as controversial that were initially labelled as controversial (True Positives). To be precise there are 549,145 True Positives and 1,899 False Negatives at this point, which shows that the classifier was able to only classify a small number of pages as controversial that were labelled as non-controversial. which could

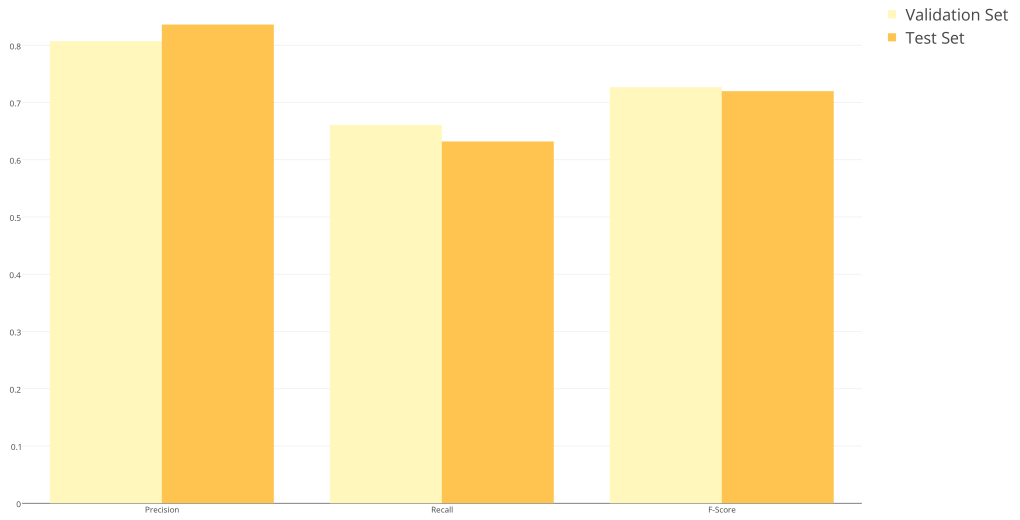

Figure 5.2: Precision, Recall and F-Score Result for Validation and Test set

also be as a result of the classifier not having sufficient information at that point to classify the pages. It is also important to note that at this time, the value of the Information Gain for all the features is zero. The Recall however, decreases to 0.3 at the 0.1 day age segment and starts to increase as the age of the page increases up until the 2.5 days when the rate of increase is almost zero.

The precision on the other hand, at the 0 day age segment is about 0.5 and starts to increase until a peak period when the precision starts to decrease. This period is the highest period among all the periods considered, when the classifier predicts the largest amount pages to be controversial that are labelled controversial. That is, at this period the ratio of the number of pages predicted as non-controversial but were initially labelled as controversial (False Positives) to the number of pages classified as controversial that were initially labelled as controversial (True Positives) is the highest compared to other periods. The peak period in this experiment for the test data is 0.3 days, which is 7 hours and 12 minutes after the page has been created. The rate at which the Precision decrease starts to reduce to almost to zero from the 1.5 day.

Taking a further look at the metrics at this peak period, Table 5.5 shows the evaluation result of the Time Dependent Experiment at this point and Table

Table 5.3: Percentage of pages correctly classified for the 10 most frequent Subreddit from all pages from which the controversial topics are extracted from

Subreddit	PCC(%)
AllSubreddits	75.41
AskReddit	76.67
leagueoflegends	75.92
nba	75.28
soccer	78.04
funny	82.51
worldnews	76.80
DotA2	77.45
nfl	83.31
SquaredCircle	76.29
news	79.89

5.6 shows the ranking of the Information Gain of all the features at this point. From the ranking of the Information Gain of the features, the *maximumNumOfCommentsInTimeSequences* is the highest ranked feature among all the features. This is because this feature takes advantage of the Time Segment's Comments and had great impact in the classification at the early stages of the pages.

Hybrid Experiment

The hybrid experiment combines the Time Independent Training set and Time Dependent Test set. Therefore the classifier is built using the the Training set that has Pages with all its related Comments, while predictions are made on the Test set that has Pages with all the Age Segment's Comments at that time segment. The purpose of this test is find out the impact of having the full information in the training set that is used in building classifier and making predictions on the test set at each Age Segment. The Precision, Recall, F-Score and the Percentage of Pages Correctly Classified metrics are also used to evaluate this experiment

Although, the results of the Hybrid Experiment as illustrated in Figure 5.5, looks in general like that of the Time Dependent Experiment, there are some major difference that are worthy of note. Comparing the Recall and the Precision for these two experiments, Figure 5.6 shows that for the Recall on the

Table 5.4: Time Independent Experiment Features Information Gain Ranking

Features	Family	Info Gain
numOfComments	Structural	0.422
maximumNumOfCommentsInTimeSequences	Age Dependent	0.399
depth	Structural	0.387
controversialMix	Sentiment	0.371
percentageOfReplies	Structural	0.346
neutralFraction	Sentiment	0.335
controversialAuthorsInPagePercentage	Strucrural	0.335
negativeFraction	Sentiment	0.325
percentageOfQuestions	Linguistic	0.325
totalNumberOfWords	Structural	0.323
percentageOfDeletedAuthors	Structural	0.322
positiveFraction	Sentiment	0.317
avgNumOfControversialContributorsInPage	Structural	0.267
wordProportion	Lingustic	0.264
avgNumOfContributorsInPage	Structural	0.262
percentageOfBadWords	Lingustic	0.228
percentageOfAdverbs	Lingustic	0.221
contradictionScore	Sentiment	0.216
percentageOfIsInPastTense	Lingustic	0.215
percentageOfModals	Lingustics	0.215
percentageOfPunctuations	Lingustic	0.212
stdDevTimestamp	Structural	0.211
percentageOfPersonalPronouns	Linguistic	0.210
percentageOfVerbs	Linguistic	0.209
percentageOfNouns	Linguistic	0.201
isSubmissionByControversialAuthor	Structural	0.075
commentsPerAge	Age Dependent	0.063
percentageOfSarcastics	Linguistic	0.026
averageTimestamp	Structural	0.016
percentageOfGilded	Structural	0.009
isAuthorDeleted	Structural	0.009
numOfGildes	Structural	0.001

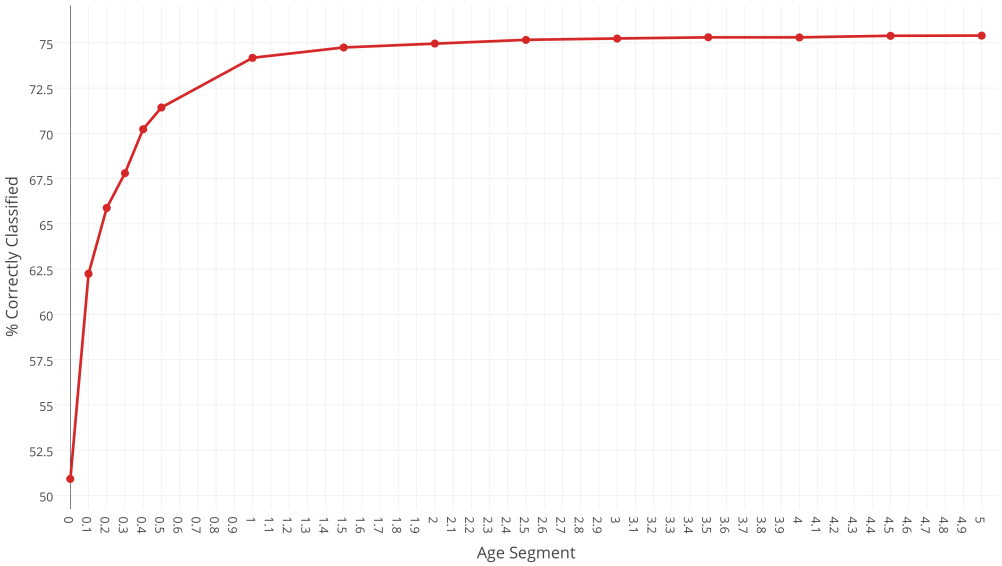

Figure 5.3: Percentage of pages correctly classified for Time Dependent Experiment

0 day age segment, while the Recall is almost one for the Time Dependent Experiment which is the highest Recall for all the time segments and reduces to the lowest at the 0.1 day age segment, the Recall for the Hybrid Experiment behaves in an opposite manner. The Recall is almost zero at the 0 day age segment and increases to the highest at 0.1 day age segment. This means that unlike what happens in the Time Dependent Experiment at the 0 day age segment in which False Negative is much lower than the True Positives, in the Hybrid Experiment at this time, the True Positives are much lower than the False Negatives. This is because the classifier is trained using a set containing sufficient information while the prediction was done on a set with insufficient information about the pages in the set.

The Precision for the Hybrid Experiment also behaves similarly to the way the Time Dependent Precision behaves, but the Precision was at the highest at 0.5 day age segment in the Hybrid Experiment, unlike in the Time Dependent Experiment when the highest Precision was at the 0.3 day age segment. Table 5.7 shows the evaluation result at this time segment. Furthermore, it is also important to note that the Precision for the Hybrid Experiment is lower than the Precision for the Time Dependent Experiment at each age segment in the experiments.

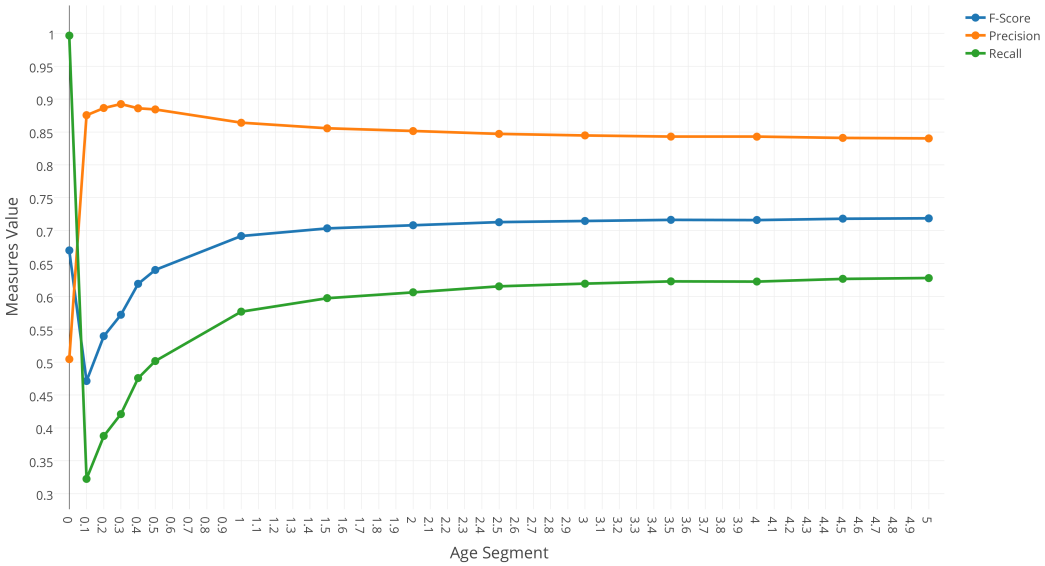

Figure 5.4: Precision, Recall and F-Score Result for Time Dependent Experiment

Furthermore, Figure 5.7 compares the percentage of pages classified correctly in both the Time Dependent and Hybrid Experiments. This shows that the percentage of pages classified correctly for the Time Dependent Experiment at the beginning is lower than that of the Hybrid Experiment. This trend continues until the 2 days age segment when the percentage for the two experiments are almost the same. This is the time when for the Time Dependent Experiments, there are sufficient information for the classifier to correctly classify the page as either controversial or non-controversial. Furthermore, according to Figure 4.2, there are only two comments per submissions at this age. This means that the training set as well as the test set has almost the same information after this time, as there are only a small number of comments after this time.

Table 5.5: Time Dependent Experiment Test Data Result at Precision Peak Period

Metric	Value
True Positives	230200.0
False Positives	27746.0
True Negatives	494877.0
False Negatives	316633.0
% Correctly Predicted	67.8
Precision	0.89
Recall	0.42
F-Score	0.57

5.3 Summary

In this chapter, the different experiments conducted were presented, the evaluation metrics used in this thesis were first presented and explained. These metrics enable us to evaluate the performance of the Classifier and how it behaves with different sets of data. Section 5.2 then went into the details of the experiments conducted and discuss the results from the experiments.

Table 5.6: Time Dependent Experiment Features Information Gain Ranking at Precision Peak Period

Features	Family	Info Gain
maximumNumOfCommentsInTimeSequences	Age Dependent	0.391
numOfComments	Structural	0.329
controversialMix	Sentiment	0.285
depth	Structural	0.282
percentageOfReplies	Structural	0.264
neutralFraction	Sentiment	0.262
controversialAuthorsInPagePercentage	Strucrural	0.260
negativeFraction	Sentiment	0.250
percentageOfQuestions	Linguistic	0.247
percentageOfDeletedAuthors	Structural	0.245
positiveFraction	Sentiment	0.243
totalNumberOfWords	Structural	0.232
stdDevTimestamp	Structural	0.223
avgNumOfContributorsInPage	Structural	0.208
avgNumOfControversialContributorsInPage	Structural	0.205
wordProportion	Lingustic	0.187
percentageOfAdverbs	Lingustic	0.167
percentageOfBadWords	Lingustic	0.166
percentageOfPunctuations	Lingustic	0.164
percentageOfIsInPastTense	Lingustic	0.163
percentageOfPersonalPronouns	Linguistic	0.162
percentageOfModals	Lingustics	0.162
percentageOfVerbs	Linguistic	0.159
contradictionScore	Sentiment	0.157
percentageOfNouns	Linguistic	0.156
commentsPerAge	Age Dependent	0.078
isSubmissionByControversialAuthor	Structural	0.077
averageTimestamp	Structural	0.020
percentageOfSarcastics	Linguistic	0.015
isAuthorDeleted	Structural	0.010
percentageOfGilded	Structural	0.008
numOfGildes	Structural	0.001

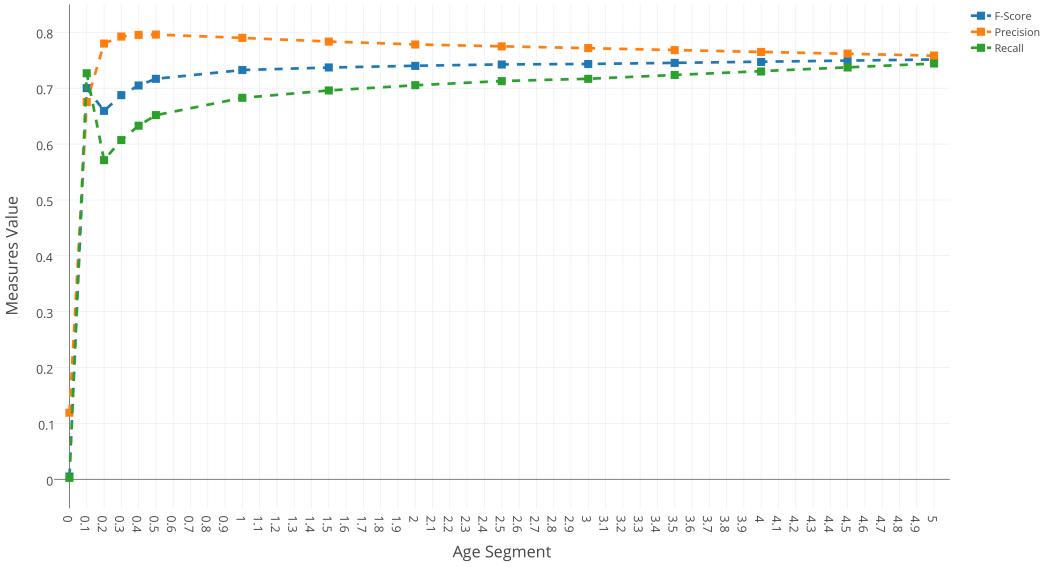

Figure 5.5: Precision, Recall and F-Score Result for Hybrid Experiment

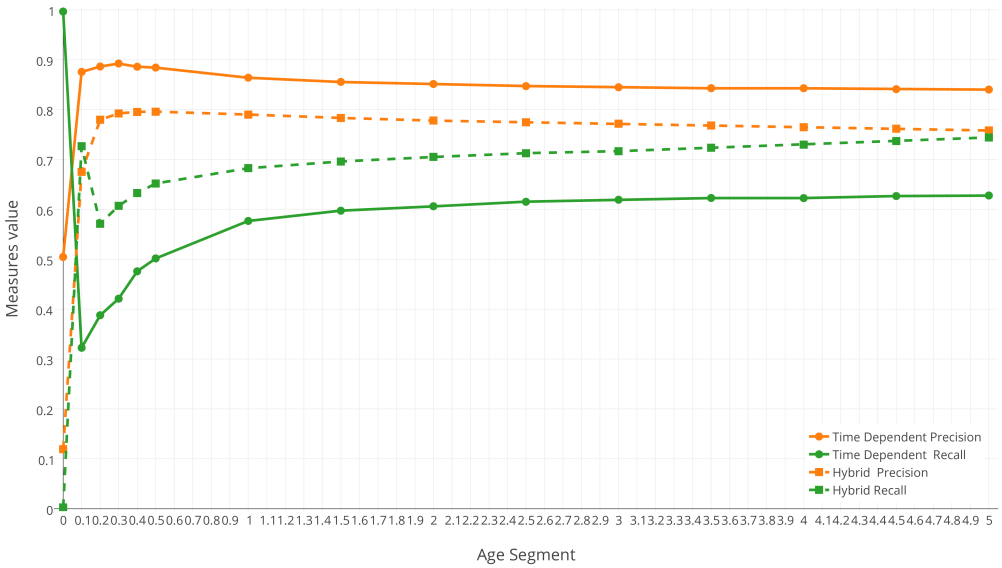

Figure 5.6: Precision, Recall and F-Score Result for Time Dependent and Hybrid Experiments

Table 5.7: Hybrid Experiment Test Data Result at Precision Peak Period

Metric	Value
True Positives	392589.0
False Positives	114192.0
True Negatives	434647.0
False Negatives	158378.0
% Correctly Predicted	75.2
Precision	0.77
Recall	0.71
F-Score	0.74

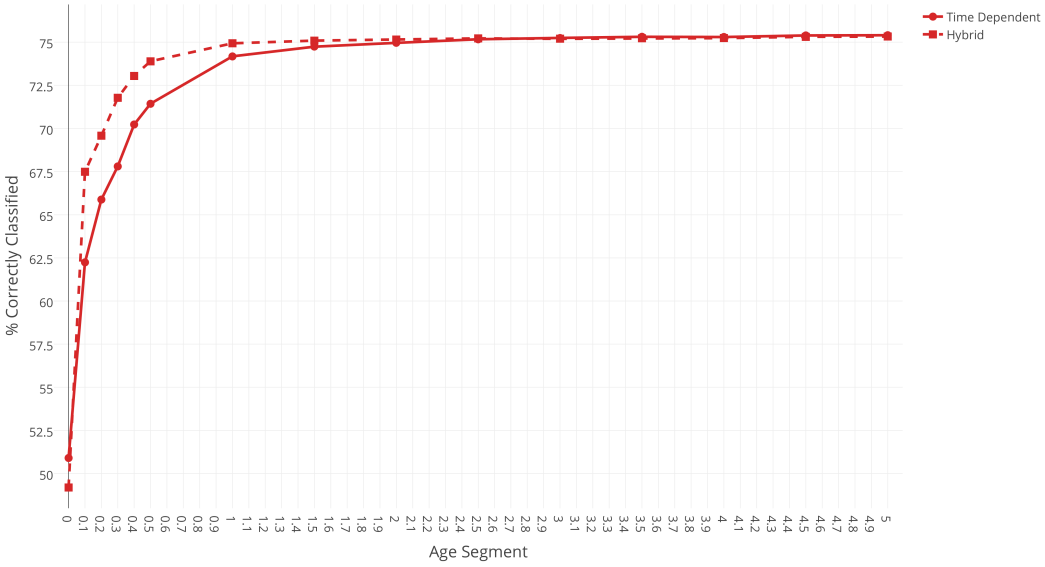

Figure 5.7: Percentage of pages correctly classified for Time Dependent and Hybrid Experiments

Chapter 6

Conclusion

We have discussed various aspects of our research in identifying controversial topics in a large-scale social media data in this thesis. In the process of our research, we created a balanced dataset that can be used in this research and made them accessible for further data analysis. Using this dataset, we develop an approach that can identify controversial topics and evaluate the performance of the approach, we also evaluate all the features used in describing the dataset and report on how efficient they are in detecting controversies. This features include some existing controversy detection measures and some created during the course of this research. In this last chapter of this thesis we shortly summarize our research and its main findings, and point out possible improvements and avenues for future research.

Main Contributions and Findings

At the beginning of this thesis, we stated some of the research questions we intend to answer in this work. One of them was the question of how difficult the task of finding controversial topics in a large dataset is. In doing this, we did a review of past research done in this area, from which we found out that there are many challenges in controversy detection, which has led to several approaches used in different research in solving this task. Using the Reddit dataset, in Chapter 3, we described the characteristics of this dataset and proceeded to extracting a subset from the whole dataset. The main contribution of this extracted subset is that firstly it provides a balanced set of controversial and non-controversial dataset that was used in our research and that can

be used in future research related to ours. Secondly, the cleaning up of the data, such as the removal of Submissions without Subreddit and Comments, provides a dataset that truly represent the real world.

In Chapter 4, the approach used in this thesis was described, and representing the extracted dataset as a Page consisting of a submission, the submission time and all the comments related to the submissions, in Section 4.2, we provide a set of pages ordered by the submission time. Furthermore, in Section 4.3, where the features used in the classification was presented, we also designed some new features apart from the known controversial measures, which can be used in future research.

From the results presented in Chapter 5, we found out that using all features presented in our work, the classifier was able to classify about 75.4% of all the pages correctly in the test set. Also from one of the points raised by Dori-Hacohen et al [DHYTA15] about the disagreement among researches on whether sentiment metric should be relevant for controversy detection or not, we found out that from the results of the ranking of the Information Gain for all the features, the four features in the Sentiment Family were among the top 20 features out of the 32 features used in this thesis, with three of them in the top 10 features. This shows that sentiment could indeed be relevant in detecting controversies.

On the second research question, which seeks to answer the question of how fast controversies can be detected, in Section 5.2 we performed an experiment called the Time Dependent, which builds the classifier and make predictions on whether a page is controversial or non-controversial based on the age of the page. We found out from the results of the percentage of page correctly classified at each time segment that at the 0 day age segment, about $50\frac{1}{2}$ of the pages are correctly classified and the percentage continues to increase to the 1 day age segment, when the rate of increase reduces significantly. This shows that by the time the page is already a day old, the classifier already has all the information needed to correctly predict a page as controversial or non-controversial. Furthermore, from the Precision metrics, we found out that from the 0 day age segment, the precision increases up until a time in which the precision starts to decrease. This point is the time when the classifier makes the most perfect prediction.

Lastly, from the third research question, in which we seek to find out how effective the known controversial measures are, only implement few of this existing controversial measures could be implemented, as most of the controversial measures found and reviewed in Chapter 2 were domain specific and not

transferable to other domains. However, out of the ones that could be adapted to our domain, from the Time Independent Experiment, taking a look at three of these controversial measures implemented, the *controversialMix* has an Information Gain of 0.371, while the Information Gain of the *wordProportion* is 0.264 and that of *contradictoryScore* is 0.216.

Future Work

In this thesis, we have described our approach in identifying controversial topics in a large-scale social media data, and have also presented the results from the various experiments conducted. In identifying controversial topics, as explained in Section 4.4, we used two simple methods in extracting controversial topics from the sets of pages classified by the classifier as controversial. However, due to the large of topics extracted: for example the first method, which uses the title of the submissions of the selected controversial pages as the controversial topics produces a total of 73,811 controversial topics, it may be of interest to further find a better way of extracting topics from these selected controversial pages.

In addition to this, in Section 4.4 we pointed out that for the Time Dependent Experiments, the *maximumNumOfCommentsInTimeSequences* features, which belongs to the Age Dependent Family has the highest Information Gain. Therefore, further work can be done in developing more features that belong to this family and to conduct experiments to find out their effectiveness.

Also in presenting the result, we examined the features in the Sentiment Family and pointed out that these features performed well in detecting controversies. However, a further survey can be done on each of the families, by carrying out experiments using only the features in the family to find out their performance in detecting controversies.

List of Figures

3.1 Cumulative frequency distribution of the number of controversial comments per submissions 21

3.2 Image Illustrating the Dataset Extraction 27

4.1 Overview of our Approach 30

4.2 Rates of Comments per Time Segment 31

5.1 Positives Negatives Metric Result for Validation and Test set . . 46

5.2 Precision, Recall and F-Score Result for Validation and Test set 47

5.3 Percentage of pages correctly classified for Time Dependent Experiment 50

5.4 Precision, Recall and F-Score Result for Time Dependent Experiment 51

5.5 Precision, Recall and F-Score Result for Hybrid Experiment . . 54

5.6 Precision, Recall and F-Score Result for Time Dependent and Hybrid Experiments 54

5.7 Percentage of pages correctly classified for Time Dependent and Hybrid Experiments 55

List of Tables

1.1 Table showing different approaches with examples of where they were used. 4

2.1 Feature List used by [PP10] 13

2.2 Participants description of Reddit 16

2.3 Participants use of Reddit features. 16

3.1 Selected Fields from the Dataset Schema 19

3.2 Important Statistics on Whole Dataset 20

3.3 List of Top 20 Subreddit with the highest number of Submissions 22

3.4 Top 10 submissions with the highest number of Comments and their Subreddit 23

3.5 Top 15 Subreddit with the highest number of Comments 24

3.6 Important Statistics from Extracted Dataset 25

3.7 Top 20 Subreddits and Time Pairs 25

3.8 Representation of the number of submissions in extracted dataset 26

4.1 Overview of the Features used in our work 37

4.2 Time Frame of the Dataset Divisions 38

4.3 Overview of the number of pages in the dataset 38

5.1 Time Independent Experiment Validation Data Result 44

5.2 Time Independent Experiment Test Data Result 45

5.3 Percentage of pages correctly classified for the 10 most frequent Subreddit from all pages from which the controversial topics are extracted from 48

5.4 Time Independent Experiment Features Information Gain Ranking 49

5.5 Time Dependent Experiment Test Data Result at Precision Peak Period 52

5.6 Time Dependent Experiment Features Information Gain Ranking at Precision Peak Period 53

5.7 Hybrid Experiment Test Data Result at Precision Peak Period . 55

Bibliography

[BES10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). 4.2

[BKSV15] Siarhei Bykau, Flip Korn, Divesh Srivastava, and Yannis Velegrakis. Fine-grained controversy detection in wikipedia. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1573–1584, 2015. 1, 1.1, 2.2, 2.2

[DHA13] Shiri Dori-Hacohen and James Allan. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 1845–1848, New York, NY, USA, 2013. ACM. 2.1

[DHA15] Shiri Dori-Hacohen and James Allan. *Advances in Information Retrieval*. Springer International Publishing, 2015. 2.1

[DHYTA15] Shiri Dori-Hacohen, Elad Yom-Tov, and James Allan. Navigating controversy as a complex search task. In *ECIR'15*, 2015. Electronic proceedings only. 2.1, 6

[ES06] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LRECâ06*, pages 417–422, 2006. 4.2

[Fel98] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998. 4.2

[Gil13] Eric Gilbert. Widespread underprovision on reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 803–808, New York, NY, USA, 2013. ACM. 2.3

[HFH+09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. 4.4

[HTF08] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer International Publishing, 2 edition, 2008. 2.1

[Kar93] Kenneth L. Karst. *Law's Promise, Law's Expression*. Yale University Press, 1 edition, June 1993. (document)

[KSPC07] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 453–462, New York, NY, USA, 2007. ACM. 2.1, 2.2, 2.2

[Mil95] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. 4.2

[Mye86] Eugene W. Myers. *An O(ND) difference algorithm and its variations*. Springer International Publishing, 1 edition, 1986. 1

[PP10] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1873–1876, New York, NY, USA, 2010. ACM. 1.1, 2.2, 2.1, 4.3, 6

[RB12] Hoda Sepehri Rad and Denilson Barbosa. Identifying controversial articles in wikipedia: A comparative study. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 7:1–7:10, New York, NY, USA, 2012. ACM. 2.2

[SCRJ04] Susan Spencer, Margo Collins, J. Albert Rivero, and Geroge Justice. *The eighteenth-century novel*. AMS Press, 4 edition, 2004. (document)

[SFM+14] Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. Evolution of reddit: From the front page of the internet to a self-referential community? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 517–522, New York, NY, USA, 2014. ACM. 2.3

[TPD] Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. Scalable detection of sentiment-based contradictions. *CEUR-WS*. 2.1, 4.3

[VLS+08] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw, and Kuiyu Chang. On ranking controversies in wikipedia: Models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 171–182, New York, NY, USA, 2008. ACM. 1, 1.1, 2.2, 2.2

[YP97] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. 5.1

[YSR+12] Taha Yasseri, Robert Sumi, Andras Rung, Andras Kornai, and Janos Kertesz. Dynamics of conflicts in wikipedia. *PLoS ONE*, June 2012. 2.1