IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. I (July – Aug. 2015), PP 17-28
www.iosrjournals.org

Analysis of Titles from the Questions of the Stack Overflow Community Using Natural Language Processing (NLP) Techniques

Tapan Kumar Hazra[1], Aryak Sengupta[1] and Anirban Ghosh[1]
[1]*Department of Information Technology Institute of Engineering & Management*
Salt Lake Electronics Complex, Kolkata-700091,West Bengal, India

Abstract: *Major online "Question and Answer" forums have proven to be of enormous help to programmers and developers from all parts of the world. One such important forum is the Stack Overflow community. In this paper, we explore and analyze the title of a question posted on the Stack Overflow community and check whether the title abides by the set of rules and guidelines defined by the Stack Overflow community [1] and [3]. We also carry out sentiment analysis on the title to judge the virality [2] quotient of the question. We present an application (or tool) developed using the Natural Language Toolkit (NLTK) and Py-stackexchange API (Application Programming Interface) of Stack Overflow in Python. Using this application, we determine the sentiment of a given title and also judge whether the title is properly formatted or not. These findings are taken into account to check whether the popularity of the question is affected by the format and sentiment of the title.*

Keywords: *NLP, Stack Overflow, Sentiment Analysis, Python, Virality.*

I. INTRODUCTION

This paper focuses on the analysis of titles of the questions from Stack Overflow community. The titles are analyzed and judged on the format and also on the sentiment. This paper tries to find a solution to the endless debate about the format of the title and also checks and determines whether the sentiment of the title plays a key role in the reach and effectiveness of the question.

The Stack Overflow community is the largest community among the major Question-and-answer websites and comprises of wide range of "topics" or "tags" related to programming and development such as Java, Python, JavaScript, Android Development, iOS Development, VHDL and many more. Answers are received in 10-15 minutes for questions which fall under the above mentioned "tags" or popular "tags" The above mentioned advantages might be the sole reason for the incredible popularity of Stack Overflow, but there are downsides to it as well. Stack Overflow comes with a very strict Voting mechanism which becomes even more difficult for a beginner to handle. The up-votes and down-votes received by a user reflect in the overall reputation of the user. Questions and answers which are written in a very uncanny or strange way result into down-votes, which even decrement the overall reputation of the user. The major reasons behind receiving down-votes are as follows:

- Questions showing no **Research** effort.
- Endeavors framed but not mentioned in the question. **"What have you tried?"** is a very common reply to questions which does not consist of personal effort.
- Questions consisting of **broken** links are likely to receive **down-votes**.
- Titles of questions consisting of negative polarity or negativity in their posts are also contenders of receiving down-votes. [2]
- If the title of the question does not abide by the rules mentioned in [1] and [3], then the question is ill-formatted and is a contender of receiving down-votes

Out of the above mentioned reasons, we mainly focus on the indented ones in this paper.
The following is the **snapshot** of a question which consists of a poorly formatted title. Now, the questions are why the title is poorly formatted. Here the title starts with "**How do I**" which is a poor format as mentioned in [1]. Though the **title does not account solely** towards the heavy down-voting of the question, but the title is definitely one of the reasons for the question being down-voted. The question also has readability issues [4] and is very poorly written which also makes the question a candidate for down-voting. The **down-votes (-11)** can be clearly seen on the left-side below the title of the question.

Fig 1: An example of a poorly formatted title (begins with "**How do I**")

Stack Overflow has a set of vocabulary which we describe as follows:
Users - A user is a registered member of the Stack Overflow web site. A user can comment, up-vote, down-vote, ask a question or even answer a question in the Stack Overflow community.

Reputation – The reputation of a user is the individual score of the user which he/she has earned in the Stack Overflow community by asking and answering questions. This score also provides certain amount of privileges to the users (e.g. – A user with above 50 reputations has the right to write a comment).

Accepted Answer – A user who has asked a question has the ability to choose an answer as the "Accepted" one among all the answers which have been written for that question.

Votes – Votes can be of two major types, Up-votes and Down-votes and votes play a major role in the reputation of a user.

Tags – When a user asks a question, he/she is required to add relevant "**tags"** to the question. The quality of the question depends on the kind of tags added to the question.

Title – Each question must consist of a **title** which must **summarize the question in one line** and **should be well-formatted to gain attention** of fellow users for getting quality answers. Hence, importance of a well-formatted **title is very essential** in a community like Stack Overflow.
Seasoned developers and programmers frame their questions by keeping in mind the above mentioned points but beginners and newcomers in the world of **Stack Overflow** find it very difficult to do so and hence, become certain contenders for receiving **down-votes**.

We believe the title plays a key role in the framing of a good question. Our paper focuses mainly on two important issues which involve the title of the question. In **Section-1 of the application**, we analyse closely whether the title abides the rules and regulations of [1]. In this section, we use the **Stack Exchange Data Explorer** to extract Titles from the questions and then also extract the tags for the given question. Here the title is verified against the format discussed in [1].

II. PROBLEM STATEMENT

In this paper, we mainly focus on two important questions which are described in the subsequent sections. The goal of this paper is to find the importance of these two issues and determine whether these issues at all are significant for the construction of a good question.
Q1 – Does high-rated (high scoring) questions consists of titles which abide by the rules discussed in [1] and [3]?
Q2 – Does sentiment of the title of a question play a significant role in the success of a question? Do titles which consist of positive sentiment draw more attention? Is the virality theory given by Jonah Berger and Katherine L. Milkman true for Stack Overflow question-titles [2]?
Q3 – Which combination of sentiment and format of the titles fetch what results? For example, a possible example will be, "Titles carrying positive sentiment but are poorly formatted" etc.

The answers to the above mentioned question can only be given if an application is developed for checking the above mentioned issues. This application is described in two sections, namely Section – I and Section – II.

III. RELATED AND EXISTING WORK

Stack Overflow is a very popular Q-and-A website which has a very strong community for programming related questions and answers. It comprises of tags ranging from "Java" to "Android Development". Apart from these advantages, there is one downside to it**. Stack Overflow** has with a very strict and tough **Voting** mechanism which becomes very difficult for a user to handle especially for a beginner. The **up-votes** and **down-votes** received by a user reflect in the overall **reputation** of the user. A drop or in reputation may also lead to blocking of that user. Our main objective in this paper would be to analyze and judge the role of creating or writing a well-formatted title for a question which minimizes the chances of getting down-votes.

There are some related works which also takes into consideration the voting mechanism. One such work is done in **"How the Stack Overflow Community Creates Quality Postings" [4]** where the researchers have analyzed the importance of **source-code** in both questions and answers of Stack Overflow. Their study revolves around the question that whether the presence of **Source-code** in questions and answers results in up-votes or down-votes of questions and answers. Their study **[4]** revealed **that a question which consists of large amount of source-codes is unlikely to do well** in Stack Overflow, whereas **answers (accepted ones) score higher when they have more source-codes attached with them.**

Apart from this, **[4]** also concludes about the readability and sentiment of a question. The readability of a particular question or answer is given by **Flesch-Kincaid Reading Ease (FKRE).** In **FKRE lower scores are more difficult** to read than higher scores. Hence, questions or answers which have higher FKRE score have greater readability than questions or answer which have lower readability scores.

Apart from the readability, another important observation in **[4]** is that questions or answers **carrying positive sentiment are more likely to receive more up-votes than questions or answers which carry neutral or negative sentiment**. Although, there are exceptions, where questions or answers with negative sentiment have high scores, but questions or answers with positive words like **"excellent", "brilliant", "enjoyable"** etc. are likely to have higher scores than other questions and answers who do not have such words with positive sentiment.

This paper by **Jonah Berger and Katherine L. Milkman [2]** on **virality** of online-content is determined by analyzing the sentiment of the content. Content that **evokes high-arousal** positive (awe) or negative (anger or anxiety) emotions is more viral. Content that **evokes low-arousal**, or deactivating, emotions (e.g., sadness) is less viral. We apply the same principle for determining and checking the title of a question for virality. Hence, we try to establish that if a title has positive content, it is likely to have more up-votes (being more viral).

This paper **[5]** explores the utility of attitude types for improving **Q-and-A** on web based communities like **Stack Overflow**. In **[5],** the researchers use attitude annotations and create a classifier for determining two types of attitudes – **argument and sentiment**.

Another research work, targeting similar issues is **[6]** where the researchers have studied and analyzed the **closed** questions in Stack Overflow. A question can be '**closed'** for five reasons – **duplicate, off-topic, subjective, not a real question and too localized**. Their key finding was to efficiently determine that many closed questions were high on information gain and was very popular among the users.

There is not much of existing work available which focus on the title of a question (to the best of our knowledge). Hence, our proposed approach for determining whether the Title is well-formatted or not is given below.

IV. PROPOSED WORK AND APPROACH

Section-1 **–** The first section determines the whether the title follows the rules mentioned and discussed in [1] and [3]. The application checks the title by validating against the rules which are discussed in [1]. Firstly, we extract the title of the question using the **Py-stackexchange API** and we also extract the tags mentioned in the question. Now, we validate against the rules mentioned in [1] and secondly, we determine whether the tags are mentioned in title of the question. If the tags are mentioned in the title of the question, then the title is ill-formatted. This assumption is also derived from [1]. Hence, we conclude from the above findings, whether the title is in accordance with Jeff Atwood's rules mentioned in [1].

Section-2**–**This section focuses entirely on the determination of sentiment from the title of the question. In this case, we follow the same approach as in **Section-1**for extracting the title using **Py-**

stackexchange API. We have used a Naïve Bayes Classifier as a classifier for carrying out sentiment analysis. The description for **Naïve Bayes Classifier** is given below. It uses the Naïve Bayes theorem.

Naive Bayes Formula [8] -
P (label | features) = P (label) * P (features | label) / P (features)

P (label) is the prior probability of the label occurring, which is the same as the likelihood that a random feature set will have the label. This is based on the number of training instances with the label compared to the total number of training instances. For example, if 60/100 training instances have the label, the prior probability of the label is 60 percent.

P (features | label) is the prior probability of a given feature set being classified as that label. This is based on which features have occurred with each label in the training data.

P (features) is the prior probability of a given feature set occurring. This is the likelihood of a random feature set being the same as the given feature set, and is based on the observed feature sets in the training data. For example, if the given feature set occurs twice in 100 training instances, the prior probability is 2 percent.

P (label | features) tells us the probability that the given features should have that label. If this value is high, then we can be reasonably confident that the label is correct for the given features.

The above description shows how the Naïve Bayes Classifier works and classifies against a test-data. Now, in order to achieve higher accuracy rate (as high as 81 %) we incorporate some techniques

Sentiment Analysis **–** The term **Sentiment Analysis or opinion mining** is mainly a technique to identify what kind of sentiment a text carries. The definition from Wikipedia **[7]** is given below:

"**Sentiment analysis** (also known as **opinion mining**) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader)."

From the above mentioned definition, it is quite clear that Sentiment Analysis aims to determine and classify a particular text or document based on its polarity. There can be two types of classifiers, namely, a binary classifier or a multilabel classifier. A **binary classifier has two labels** – **positive and negative** while a **multilabel classifier can assign one or more labels** to a piece of text.

Corpus – We have used the **movie reviews corpus [8]** where the movie reviews are categorized and classified in **positive and negative** sentiments**.** We have **trained** our **Naïve Bayes Classifier** using the movie reviews corpus **[8].**

Now, the accuracy rate of the classifier depends mainly on the scheme via which feature extraction is done. **Feature extraction** is the process via which an input data is transformed into a set of features for the sake of avoiding **redundancy**. First we extract features using the **Bag-of-words** feature extraction technique. Training the classifier with this set of features, we acquire an efficiency of **72.8 %.** This accuracy can be further improved by two techniques which are mentioned below:

- Removing or filtering out **Stop words**. The stop words are words like **"the", "an", "but"** etc.
- Including **bigrams or collocations** result in a significant change in the accuracy of the classifier.

Firstly, we filtered out stop words and found that there is a **decline in accuracy of 0.2 %** since, the accuracy dropped from **72.8 % to 72.6 %**. This decline in accuracy shows that stop words add some valuable information for classification despite being a **noise** in the **data set**.

Secondly, we included bigrams into the feature set and we marked a significant incline in accuracy of **9%,** i.e. the accuracy increased from **72.8 % to 81.6 %.**

Hence, we conclude that inclusion of bigrams result in a significant amount of increase in the accuracy level for a classifier.

V. APPLICATION

The tasks or jobs mentioned in Section-1 and Section-2 are carried out with the help of a **Python** script or a python powered application. The application first determines and concludes whether a title of a question is in accordance with the rules posted in **[1] and [3].** If the title is properly formatted, then we conclude the same. If the title is not properly formatted then we conclude that the title is not properly formatted with the exact reason behind the title not getting formatted.

After this, we directly check and analyze the sentiment of the title with the techniques mentioned in **Section-2.** The sentiment is judged and printed accordingly, based on which we provide a final verdict for the title of the question.

The question is heavily down-voted and also consists of tags in the title of the question. The application was able to detect this (presence of tags in title) with the help of **Py-stackexchange API, NLTK Toolkit and the Pattern.en module**.

This application takes the **Question ID** as input and checks and determines everything mentioned in Section -1 and Section -2.

***Algorithm*:**

The algorithm used for the project is given below:

Step 1: The entire dataset of Stack overflow site, namely DataSetFinal, is fetched in order with following fields Id, **Title**, **Tags**, **Score**, **AnswerCount**, **FavouriteCount**, **ViewCount**, **CreationDate and ClosedDate**. DataSetFinal sheet is read using csv module.

Step 2: Titles of each question are checked whether titles are properly formatted or not using '**titlecheck()'** function.

Step 3: Each title is tokenized into list using the object of RegexpTokenizer module of **nltk** library and respective tags are also converted into list which was strings previously to check whether tag words exist in title or not in '**titlecheck()'** function.

Step 4: Each tokenized title is checked with respect to pre-defined titles (Title starts with "How do I", "How can I", "How do you") which are signs of bad format and type of title format and reason behind that are set accordingly in '**titlecheck()'** function.

Step 5: Sentiment of each title is analysed using **sentiment** function of **pattern.en** module. This **sentiment** function returns polarity and subjectivity which are used to decide the type of sentiment.

Step 6: If ((subjectivity == 0.5 and polarity == 0.5) or polarity == 0.0 or (subjectivity>polarity and polarity>0.0)) holds true then sentiment is set to **neutral** and if (polarity<0.0) holds true, sentiment is set to **negative** otherwise NaiveBayesClassifier is used to find sentiment for the else (or positive) part.

Step 7: The NaiveBayesClassifier classifier is trained first with movie_review corpus of the nltk toolkit.

Step 8: Title is tokenized into list using the object of RegexpTokenizer module of **nltk** toolkit and best_words_and_filtered_stop_words_feats function is called to filter out all stop words. best_words_and_filtered_stop_words_feats() function returns a dictionary where each title word is set to true or false value.

Step 9: Filtered title is classified with trained classifier which returns the type of sentiment of the title.

Step 10 : A excel sheet ,namely FinalResultsSheet, is generated using **xlwt** module with following fields Id, Title, Title Format, Title Remarks, Sentiment, Score, Answer Count, Favourite Count, View count, Creation date, Closed Date.

Step 11: Excel sheet is written automatically when application is run.

Step 12: Several statistics are calculated using FinalResultsSheet excel sheet.

VI. DATASET DESCRIPTION

Stack Overflow provides all user-generated content on its website for download under the Creative Commons Attribute-ShareAlike license. We have downloaded a part of the Dataset from the Stack Overflow community using the **Stack Exchange Data Explorer [11]** from the Stack Exchange data dump provided by Stack Overflow community. This copy of the actual data set is updated weekly with new set of "Posts" and User specific data. The Stack Exchange Data Explorer [11] provides a interface to run SQL commands and fetch and download a part of the data dump from the Stack Overflow community in a "**csv**" (Excel sheet) format.

VII. ANALYSIS

In this paper, our study mainly revolves around the importance of titles for a question in the Stack Overflow community. We make a thorough analysis of the dataset to judge whether the format of the title or the sentiment which the title caries contributes to the success rate of the entire question. We use the Python powered application to determine and verify the effectiveness of the titles and also comment on the three problem statements which are described and defined previously.

Before answering or commenting on the two research questions or problem statements **[Q1], [Q2] and [Q3]** we do some groundwork which will lead us to the actual discussion. There are a total of **22,251**(out of 50,000) active (questions which are not deleted yet) questions from the reduced dataset which we derived earlier by using the **Stack Exchange Data Explorer [11].** The following are the findings for the problem statements mentioned in **[Q1]**, **[Q2] and [Q3]**.

A. ***Findings for Q1*** **-**

The first research question or problem statement **[Q1]** states that "**Does high-rated (high scoring) questions consists of titles which abide by the rules discussed in [1] and [3]?**"

To comment on **[Q1],** we first calculate the percentage of titles which are poorly formatted or which do not abide by the rules discussed in **[1]** and **[3]**.

We have found that around **31.47% of the titles** are not properly formatted. Out of this, only **24.41%** of the questions have **positive scores,** i.e. they have been **up-voted**. Around **9.09%** of the questions have been **down-voted** or they have negative **"scores** "and around **66.49%** of the questions have a **score of zero** (i.e. they are **neither up-voted nor down-voted**).

After calculating this, we have also calculated the above statistics for questions which have properly formatted titles. Around **68.51%** of the questions consist of a properly formatted title. Out of this 68.51%, around **24.80%** have **positive scores**, i.e. these questions are **up-voted**. We also find that around **7.93%** of the questions have **negative scores**, i.e. they are **down-voted** and **67.25%** of the questions have a **score** of **zero (i.e. they neither up-voted nor down-voted).**

From the above deductions we can conclude that the format of the title of the questions does not impact drastically on the "**scores**" of the question. A slight decrease in percentage (from **9.09% for badly formatted to 7.93% for well-formatted titles**) was marked for the **down-voted** questions. Thus, from this we can conclude that, having a well-formatted title does not necessarily guarantee a question to be **up-voted** or have **positive scores**. Similarly, a poorly formatted title is also not always (only 9.09%) subjected to **negative scores** or **down-voting**. This percentage is also very close to the percentage (7.93%) of **down-voted** questions which have well-formatted titles.

There are other attributes which also contribute as a parameter for measuring **success** or **reach** of a question. These attributes are **"AnswerCount**", "**FavouriteCount**", **"ViewCount**" and whether the Question has been **"closed**" or not. We have considered the above mentioned attributes for analyzing the "**reach**" or "**traction**" gained by the poorly **([1] and [3])** formatted **titles**.

We start by analyzing the attribute "**Scores**". This attribute "**Scores**" refer to the **up-votes** and **down-votes** earned by a particular post. For our study, we are only considering the "**Scores**" of questions and not answers. We

Now we analyze the attribute "**AnswerCount**". We find that none (**28.72%**) of the poorly formatted titles has an AnswerCount of **"0"**. The average number of answers (or **average AnswerCount**) for poorly formatted titles is **1.05**. This means, for every question which consists of a poorly formatted title, will have at least **1 to 2**(**1.05**) answers on an average. Hence, from this we can also conclude that an improper title format does not affect the **AnswerCount** significantly. Even without a properly formatted title, a question has a possibility of getting either **one or two** answers on average. Now, we again run the tests for properly formatted titles and we find that a similar average of **(1.05),** i.e**.,** again a range of one to two answers. Hence, we conclude that the format of the titles is not quite significant in terms of "**AnswerCount",** since both **formatted** and **un-formatted** titles result in similar average "**AnswerCount".**

"**FavouriteCount**" is another attribute which also contributes to the popularization of the questions. We find that a badly ([1] and [3]) formatted title consists of an **average** "**FavouriteCount** "of **0.868**(close to 1). Hence, a badly formatted title has an average "**FavouriteCount"** of 0.868 (little less than 1).Now, we again run our tests for calculating the average "**FavouriteCount**" for questions which consists of properly formatted titles. The average "**FavouriteCount**" turns out to be **(0.946)** which is also very close to the 1 but again little less

than 1. Hence, we can again conclude that the format of titles does not play a key role or does not have any significant impact on the average "**FavouriteCount**".

Stack- Overflow Attributes	Format of Titles		
	Statistics Type	Badly formatted titles	Well formatted titles
Scores (+ve, -ve, zero)	Percentage	24.41 (+ve) 9.09 (-ve) 66.49 (0)	24.80 (+ve) 7.93 (-ve) 67.25 (0)
Answer Count	Average	1.05	1.05
Favourite Count	Average	0.868	0.946
View Count	Average	41.76	42.80
Closed	Percentage	3.11	2.75

Table I: Format of titles vs. stack overflow attributes

Similarly, "**ViewCount**" is another attribute which speaks volume about a question's "**popularity**" or "**reach**". We again take a similar approach as we have done for the above cases. We find that the average "**ViewCount**" for poorly formatted titles turns out to be **41.76 (i.e. around 42 views on average)**. Now again, we run the same test for **properly formatted** titles and we find that an average of **42.80 (again close to 42 but marginally more)** views are there. From this we can clearly state that the attribute "**ViewCount**" doesn't quite depend on the format of the titles. Although, a slight increase in the "**ViewCount**" is observed for well formatted titles, but that cannot be considered as a significant impact.

Whether the question is "**closed"** or not is something which we need to consider before drawing a final conclusion on the impact of the format of titles over the popularity of the questions. As we have calculated earlier, only **31.47% of the questions** have titles which are not properly formatted.

Out of this 31.47%, a percentage of **3.11%** of the questions have been **closed** by the moderators due to various reasons. Now, we consider the titles of the questions which are properly formatted and which have been closed due to various reasons, we get a percentage of **2.75%** of the total number of well formatted titles. Thus we observe a marked decrease in the percentage of closed questions for well formatted titles over poorly formatted titles. Hence, we can conclude that the questions which have badly formatted titles are more prone to get "**closed**" than the questions which have well formatted titles.

From the above mentioned statistics, we can finally comment and provide an answer for **[Q1]**. The following are the major highlights or conclusions of the above mentioned findings **[Table I]** and are again stated below:

- Having a well-formatted title does not necessarily guarantee a question to be **up-voted** or to have **positive scores.**
- Similarly, a poorly formatted title is also not always (only 9.09%) subjected to **negative scores** or **down-voting**. This percentage is also very close to the percentage (7.93%) of **down-voted** questions which have **well-formatted** titles.
- The format of the titles of the questions does not have a significantly large impact on the attributes like **"AnswerCount", "FavouriteCount" and "ViewCount".**
- The questions which have badly formatted titles are a little more prone (0.36% more chance) to get "**closed**" than the questions which have well formatted titles.

Thus, from these facts we can conclude that **overall** (apart from a few exceptions), the format of titles does not play a key role in defining the popularity of the questions. There are a few exceptions. The questions which have badly formatted titles are more prone to get "**closed**" than the questions which have well formatted titles. Apart from this, there does not exist, any significant impact of the format of titles on the **popularity** or **reach** of the questions.

B. ***Findings for Q2 -***

The second research question or problem statement states the following:

"Does sentiment of the title of a question play a significant role in the success of a question?

Do titles which consist of positive sentiment draw more attention?

Is the virality theory given by Jonah Berger and Katherine L. Milkman true for Stack Overflow question-titles [2]?" As stated above, the problem statement **[Q2]** consists of three major segments or parts. Our work-flow will consist of analysis of each and every segment coupled with various statistics which will substantiate the derived result.

First Part: We first consider the segment, "**Does sentiment of the title of a question play a significant role in the success of a question?**" For analyzing this, we start by calculating the percentages of titles which fall under the three categories of sentiments namely, "**positive**", "**negative**" and "**neutral**". We find that the percentage of questions which has titles carrying "**neutral**" sentiment is **88.48%**, whereas the titles of questions carrying "**positive"** and "**negative"** sentiments are **1.02%** and **10.49%** respectively. Thus, this result shows that the majority of the titles used for questions in the **Stack Overflow community** are devoid of any sentiment, i.e. they carry **no** or **neutral (i.e. neither positive nor negative)** sentiment.

Now, we again run some tests to find the impact of sentiment over the different attributes which we considered in the above section. These attributes are "**Scores**", "**AnswerCount**", "**FavouriteCount**", **"ViewCount**" and whether the question has been "**closed**" or not. We start by analyzing the attribute "**Scores**" with along with questions with titles carrying "**Neutral**" sentiment. We find that almost **24.61%** of the "**Neutral**" questions have a **positive score** or have been **up-voted**, **67.07**% have a score of **zero** (i.e. neither **up-voted nor down-voted**) and **8.33%** have a **negative score** or have been **down-voted**.
For questions with titles carrying "**Positive**" sentiment, we find that about **32.15%** have a **positive score** or have been **up-voted**, **58.14%** have a score of **zero** (i.e. neither **up-voted nor down-voted**) and **9.69%** have a **negative score** or have been **down-voted**.

Table II: Sentiment of Titles vs. Stack Overflow Attributes

Stack Overflow Attributes	Sentiment of Titles			
	Statistics Type	Positive	Negative	Neutral
Scores (+ve, -ve, zero)	Percentage	32.15,+ve 9.69 ,-ve 58.14,0	24.53,+ve 7.83, -ve 67.62, 0	24.61 ,+ve 8.33, -ve 67.07, 0
Answer Count	Average	1.12	1.03	1.05
FavouriteCount	Average	1.14	0.911	0.918
ViewCount	Average	44.65	44.74	42.18
Closed	Percentage	2.64	2.65	2.89

For questions with titles carrying "**Negative**" sentiment, we find that about **24.53%** have a **positive score** or have been **up-voted**, **67.62%** have a score of **zero** (i.e. **neither up-voted nor down-voted**) and **7.83%** have a **negative score** or have been **down-voted**.

We again run some tests for analyzing the attribute "**AnswerCount**". We find that an average of **1.12** answers is obtained for questions with a title carrying "**positive**" sentiment. For titles of questions carrying "**negative**" sentiment, the average number of answers is **1.03**and for titles of questions carrying "**neutral**" sentiment, the average number of answers is **1.05.** Hence, we can conclude that the "**AnswerCount**" remains "**very close to 1**" irrespective of the sentiment of the title of a question.

For the attribute "**FavouriteCount**", we find that the questions with titles carrying "positive" sentiment have an average "**FavouriteCount**" of **1.14**, questions with titles carrying "**negative**" sentiment have an average "**FavouriteCount**" of **0.911**and finally questions with titles carrying "**neutral**" sentiment have an average "FavouriteCount" of**0.918**. Hence, we can conclude that the average of the "**FavouriteCount**" attribute also remain "**very close to 1**" for all types of sentiment.

Again we run some tests for the attribute "**ViewCount**", we find that the questions with titles carrying "positive" sentiment have an average "**ViewCount**" of **44.65**, questions with titles carrying "**negative**" sentiment have an average "**FavouriteCount**" of **44.74**and finally questions with titles carrying "**neutral**" sentiment have an average "FavouriteCount" of **42.18**. So, we find that questions with titles carrying "**positive**" and "**negative**" sentiment have an average "**ViewCount**" of **45** and "neutral" titles have an average count of **42**.**Hence, we can conclude that the attribute "ViewCount" shows a slight improvement when presented with a question consisting of a title carrying of a sentiment (positive or negative) over questions with titles carrying neutral sentiment.**

We then run tests for analyzing the "**closed**" questions. The questions with titles having "**positive**" sentiment have a **2.64%** chance of getting "**closed**". The questions with titles having "**negative**" sentiment have a **2.65%** chance of getting "**closed**" and titles with "neutral" sentiment have a **2.89%** chance of getting "**closed**". Hence, the sentiment of the title of the question doesn't play a key role in the context of the question getting "**closed**".
From the above findings, we can conclude the following **[Table II]:**

- Around **32.15%** of the titles of the questions which carry "**Positive**" sentiment have been **up-voted** whereas in case of "**Neutral**" and "**Negative**" questions, it is only **24.61%** and **24.53%** respectively. **Thus we can say that a question which has a title carrying positive sentiment is more prone to be up-voted than the titles carrying neutral and negative sentiment.**

- Around **58.14%** of the titles which carry "**Positive**" sentiment have scores of "**zero**" whereas in case of "**Neutral**" and "**Negative**" sentiment, it is 67.07% and 67.62% respectively. **This shows that a question with a title having positive sentiment has a significantly less chance of scoring a zero than the questions of titles having neutral and negative sentiment.**

- Now we observe an unusual result which shows that only **7.83%** of the questions having titles carrying "**Negative**" sentiment has been down-voted whereas for "**Neutral**" and "**Positive**" sentiment it is **8.33%** and **9.69%. This shows that a question with a title having negative sentiment has a less chance of getting down-voted than the questions of titles having neutral and negative sentiment.**

- **The attribute "ViewCount" shows a slight improvement when presented with a question consisting of a title carrying of a sentiment (positive or negative) over questions with titles carrying neutral sentiment.**

Second Part - Now, from the above statistics and conclusions, we can answer this part of **[Q2], "Do titles which consist of positive sentiment draw more attention?"** The first part of the conclusion states that "**a question which has a title carrying positive sentiment is more prone to be up-voted than the titles carrying neutral and negative sentiment.**" This conclusion is based on the fact that almost 32.15% of the questions with titles having **positive** sentiment are up-voted while this percentage is 24.61% and 24.53% for questions with titles having **neutral** and **negative** sentiment. The second conclusion states that, **a question with a title having positive sentiment has a significantly less chance of scoring a zero than the questions of titles having neutral and negative sentiment.** So, from the mentioned points, it can be concluded that as far as the attribute "**scores**" is concerned, a title having "**positive**" sentiment does well over its "negative" and "neutral" counterparts.

Third Part- The third and final part of [Q2] is "**Is the virality theory given by Jonah Berger and Katherine L. Milkman true for Stack Overflow question-titles [2]?**" This question can also be answered from the conclusions drawn in the above paragraph. As far as **scores** are concerned, a question with a title carrying **positive** sentiment has a much higher chance of doing well than its **negative** and **neutral** counterparts. But the "**Virality Theory"** is not quite valid for the other attributes which also define a question's reach or popularity like "**AnswerCount**", "**FavouriteCount**" and "**ViewCount**". From the above findings we can clearly see that the sentiment does not play a key role for these attributes and as a result the **Virality Theory** also does not hold for these attributes.

C. ***Findings for Q3–***
To answer **[Q3]** properly, we take into consideration all possible combinations for both the **sentiment** and the **format of the title.** Tests are run against 4 different combinations of sentiment and format of a title. The results of the tests are listed below.
A.Not properly formatted along with "positive", "negative" or "neutral" sentiment of a title.
Here we are running tests to check a badly formatted title carrying positive, negative or neutral sentiments against the Stack Overflow attributes. This analysis helps us to draw a conclusion on all the combinations to answer **[Q3]** accurately.

Table III: Sentiment of Not Properly Formatted Titles Vs Stack Overflow Attributes

Stack Overflow Attributes	Sentiment of Not Properly Formatted Titles			
	Statistics Type	Positive	Negative	Neutral
Scores (+ve,-ve, zero)	Percentage	27.86,+ve 9.383, -ve 62.29,0	24.41,+ve 8.41, -ve 67.17, 0	24.38 ,+ve 9.16, -ve 66.45, 0
AnswerCount	Average	1.24	0.98	1.06
FavouriteCount	Average	1.14	0.76	0.87
ViewCount	Average	41.52	44.00	41.50
Closed	Percentage	3.27	3.44	1.41

The key findings from Table - III are as follows:

- Even with a badly formatted title, a question has a better chance of having a positive "**Score**" count if the title carries a **positive** sentiment. For **positive** titles, the percentage of questions having positive score is **27.86%** which is slightly more than **24.41%** and **24.38%** for **Negative** and **Neutral** titles respectively.

- A title with a poor format and a "**Negative**" sentiment has the highest (67.17%) chance of getting a "**Score**" of zero but, on the contrary, it also has the lowest (8.41%) chance of receiving a negative "**Score**".

- A title with a poor format and a "**Negative**" sentiment has the highest number in terms of "**ViewCount**" (**Around 44**), whereas it is close to 41 for others.

A title with poor format and carrying a "Negative" sentiment has the least chance of getting closed (1.41%) than the titles carrying "Positive" and "Negative" sentiment.

Table IV: Sentiment of Properly Formatted Titles Vs Stack Overflow Attributes

Stack Overflow Attributes	Sentiment of Properly Formatted Titles			
	Statistics Type	Positive	Negative	Neutral
Scores (+ve, -ve, zero)	Percentage	33.73,+ve 9.63 ,-ve 56.62 ,0	24.59,+ve 7.57, -ve 67.82, 0	24.72 ,+ve 7.95, -ve 67.31, 0
AnswerCount	Average	1.08	1.05	1.05
FavouriteCount	Average	1.14	0.97	0.93
ViewCount	Average	45.80	45.07	4 2.50
Closed	Percentage	2.40	2.29	2.81

B. Properly or well formatted along with "positive", "negative" or "neutral" sentiment of a title

The tests we run for this segment is similar to the above section, but the problem definition changes a bit. Here we are considering well formatted titles and also its sentiment as data for the tests. This will also help us comment on **[Q3]** and answer it more accurately.
The key findings from **Table – IV** are as follows:

- With a properly formatted title, a question has a much better chance of having a positive "**Score**" count if the title carries a **positive** sentiment. For **positive** titles, the percentage of questions having positive score is **33.73%** which is reasonably more than **24.59%** and **24.72%** for **Negative** and **Neutral** titles respectively.

- A title with a **good format** and a "**Negative**" sentiment has the highest (**67.82%)** chance of getting a "**Score**" of zero but, on the contrary, it also has the lowest (**7.57%**) chance of receiving a negative "**Score**". This conclusion is very similar to the one drawn for [**Table III**].

- A title with good format and a "**Neutral**" sentiment has the **lowest** number in terms of "**ViewCount**" (**Around 42**), whereas it is close to **45** for "**Negative**" and "**Positive**" titles.

Now, keeping in the view the above mentioned conclusions, we are in a position to answer [Q3].We find that a combination of ("Well-formatted" plus "Positive") titles and a combination of ("Badly-formatted" plus "Positive") has the highest chance of scoring a positive "score" over any other combinations.

On the other hand, a title which carries "Negative" sentiment for both (Badly formatted and Well-formatted) doesn't necessarily guarantee a negative "Score".

VIII. CONCLUSION

The analysis of the titles done in the above segment provides us with various conclusions. The conclusions are listed in each segment individually. In this segment, we highlight the conclusions which provide a direct impact on the "**aim**" or "**motive**" of our project.

We have concluded earlier that a well-formatted title comes with no guarantee of getting up-voted. On the contrary, a badly formatted title also doesn't come with any guarantee of getting down-voted. The **badly-formatted** titles do not provide any significant effect on **AnswerCount**, **FavouriteCount** and **ViewCount**. **Badly-formatted** titles have a slightly higher chance to get "**closed**".

After this we have calculated and analyzed the impact of the sentiment of the titles on the "**reach**" or "**popularity**" of the questions. We find that a title with positive **sentiment** is more prone to be **up-voted** than the titles carrying "**neutral**" or "**negative**" sentiment. Titles having positive sentiment have a lesser chance of **scoring a zero** than "**negative**" and "**neutral**" sentiment.

Some important calculations are also drawn with the mixed results. Titles with a poor format and also carrying a positive sentiment have a slightly more chance of gaining up-votes. A title with poor format and a **negative** "**Score**" has the highest chance of receiving a "**zero score**" and surprisingly the lowest chance of receiving a "**negative**" score. Titles with a **poor format** and **negative** sentiment are having a chance to receive slightly **higher number of views (ViewCount)**.

Titles with a **good format** and carrying a "**positive**" sentiment have a reasonably higher chance of getting up-votes than its "**neutral**" and "**negative**" counterparts. A title with a **good format** and carrying "**negative**" sentiment has the highest chance of scoring a "**zero**". Again another surprising fact turned out to be that a title with **good format** and "**negative**" sentiment has the least chance of receiving up-votes. **Thus from this it can be concluded that though the sentiment of a title can be considered as a important factor for receiving up-votes but the format of the title overshadows the sentiment factor in the case of down-votes**. It is also observed that the titles with a good format and positive or negative sentiment has a chance of getting **more views** than the titles with neutral sentiment.

IX. FUTURE SCOPE

In this paper, we have used Naïve Bayes Classification model to train and test our movie reviews corpus data. A better and efficient way is to use **Support Vector Machines** as the classification model. The Wikipedia definition is given below:

"In machine learning, **support vector machines** (**SVMs**, also called **support vector networks)** are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier"

The implementation for a **Support Vector Machine** model is not given by the NLTK framework of python. For implementing Support Vector Machines, we can use **scikit-learn [9]** using Python. A proper implementation as well as documentation of Support Vector Machines is given in **[10].**

A good amount of improvement on the accuracy of the statistics can also be obtained if we consider the entire **Stack Overflow data dump**. This data dump can be downloaded from the Stack Overflow website.

X. REFERENCES

[1]. How do I write a good title? http://meta.stackexchange.com/questions/10647/writing-a-good-title
[2]. Berger, Jonah, and Katherine L. Milkman. "What makes online content viral? "Journal of marketing research 49.2 (2012):192-205.[pdf]Can we prevent titles with an unnecessary tag in them? http://meta.stackexchange.com/questions/103563/can-we-prevent-titles-with-an-unnecessary-tag-in-them
[3]. Squire, Megan, and Christian Funkhouser. "" A Bit of Code": How the Stack Overflow Community Creates Quality Postings." System Sciences (HICSS), 2014 47th Hawaii International Conference on. IEEE, 2014.[pdf]

[4]. Somasundaran, Swapna, et al. "QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News." ICWSM. 2007. [pdf]
[5]. Correa, Denzil, and Ashish Sureka. "Fit or unfit: analysis and prediction of' closed questions' on stack overflow." Proceedings of the first ACM conference on Online social networks. ACM, 2013. [pdf]
[6]. Sentiment Analysis in Wikipedia http://en.wikipedia.org/wiki/Sentiment_analysis
[7]. Accessing text corpora and useful resources, NLTK, http://www.nltk.org/book/ch02.html
[8]. Scikit-learn, machine learning in Python http://scikit-learn.org/stable/
[9]. Support vector machines, Scikit-learn http://scikit-learn.org/stable/modules/svm.html
[10]. Stack Exchange Data Explorer http://data.stackexchange.com/

XI. AUTHOR'S BIBLIOGRAPHY

Tapan Kumar Hazra completed his M.E degree from Jadavpur University, Kolkata, West Bengal, India.

Since from 2003, he is working as Assistant Professor of Department of Information Technology at Institute of Engineering & Management, Salt Lake, Kolkata, West Bengal, India. His research interest includes Design and Analysis of Algorithms, Image Processing, Machine learning, Cryptography.

A. Sengupta, pursuing B.TECH degree in Information Technology at Institute of Engineering & Management,

West Bengal University of Technology, West Bengal, India and is in his Final year (Final Semester). His areas of interest for research include Natural Language Processing, Sentiment Analysis, Machine Learning and Stack Overflow.

A. Ghosh, pursuing B.TECH degree in Information Technology at Institute of Engineering & Management,

West Bengal University of Technology, West Bengal, India and is in his Final year (Final Semester). His areas of interest for research include Natural Language Processing, Sentiment Analysis and Artificial Intelligence.