Article 1

Identifying Fake Reviews with SAS (Statistical Analysis System)

SAS is a software suite that can mine, alter, manage and retrieve data from a variety of sources and perform statistical analysis on it

We are considering a text mining approach to derive a text-based definition for the analysis of fake reviews on Yelp.

**Purpose**:

Due to the constraint of memory, we are only including relevant data for our processing which includes fields such as Business ID, date, Review\_Id, text, user\_Id from NVLCReview Dataset.

A screenshot of a computer

Description automatically generated

**Text Parsing:**

In text parsing, we parsed through the whole dataset and found out the words that were irrelevant for the analysis of reviews. So, in order to process our data, text parsing pallet helped us reduce our dataset by removing those type of words.

**Text Filter:**

In text filtering, we filtered out all parts of speech except adjectives and defined synonyms for words that have the similar meaning.

**Text Cluster:**

For the whole process, we ran it multiple times till the time all the word with similar meanings formed different sets of cluster

A close up of a logo

Description automatically generated

A picture containing outdoor

Description automatically generated

**Text Topic:**

Text topic helps us explore the document collection by automatically associating terms and documents according to both discovered and user-defined topics.

**A screenshot of a cell phone

Description automatically generated**

After we analyzed with text cluster and text topic, we decided to use number of terms and number of documents by topic to determine with our fake review.

We compared the number of terms with number of documents by topics and derived if the number of terms is greater than the number of documents then these kinds of topics are fake reviews.

Article 2

**Identifying Fake Review using SAS Miner**

While analyzing the yelp website and the given dataset and conducting literature review on the effectives of Yelp’s review system, we found that their current review system consists of as many as 25 percent of apparent fake reviews which are false, inflationary and misleading. These suspicious reviews praise or criticize the services provided by restaurants, dentists, among many others and are either biased or consists of unhelpful rants. It was also observed that many businesses create apparent fake reviews groups on social media websites like Facebook, Twitter and other forged websites to get in touch with potential reviewers and share discount coupons and/or send free products in exchange for positive reviews.

Based on the above findings, we concluded that reviews that have an extremely positive tone are mostly written by reviewers who either are partial towards that specific business due to a hidden personal interest or are paid reviewers. On the other hand, reviews with a high negative tone are mostly written by reviewers who are either influenced by a competitor or paid by the competitor for the review.

Assumption:

We classify fake reviews as follows:

* Reviews with a very high positive tone
* Reviews with a very negative tone

We used the SAS provided AFIN Sentimental dataset which contains words having positive and negative tone along with weights.

Our input file consists of 100K records of reviews with data mostly from cities: Tempe and Pittsburgh.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

We considered id, text, stars in the dataset and rejected the rest of the fields like useful, funny, cool, user\_id, business\_id, etc

A screenshot of a computer

Description automatically generated

We first added the Text Parsing Node to the Review Data. We set the parts of speech as NO because it will not add any value in our current analysis. In Text Parsing, we used default SAS stop list and Stem term list.

A screenshot of a computer

Description automatically generated

After parsing the text, we added the Text Filter Node. In the Text Filter node, we changed the default option of ‘Check Spelling’ from “No” to “Yes” as we are going to use Sentimental Analysis using dataset of positive and negative tone. We considered frequency weighting as log.

Log of frequency provides more weights to less common words than very frequent words like “the” etc.

We added open office US English dictionary which we downloaded from below website.

<http://extensions.services.openoffice.org/en/project/en_US-dict> and run code in program editor as provided. And then attached that dataset in dictionary.

A screenshot of a cell phone

Description automatically generated

In Text Topics we used user topics and SAS provided dataset Sampsio, AFINNN sentiment table which contains words and weightage for positive and negative. We selected the number of multiple term topics as 5.

A screenshot of a social media post

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In topic viewer, we observed the below classification of topics. Top two classifications highlight the negative tone documents and positive tone documents. These two are the highly positive and negative documents which as per our assumption are fake reviews.

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

The above two snapshots highlight the positive and negative terms and weight for each term and number of documents associated. We will remove the negative and positive tone documents in the output file as these have been identified as fake.