**Research Paper**

**on**

**Polarized Opinion Maker**

**using Machine Learning.**

**Polarized Opinion Maker using Machine Learning**

**Authors:**

Aayushi Joshua | Harshita Kansara | Rajshree Patale | Tanmay Borde |

**Faculty:**

Prof. Devyani Kamble (Assistant Professor)

**Abstract:**

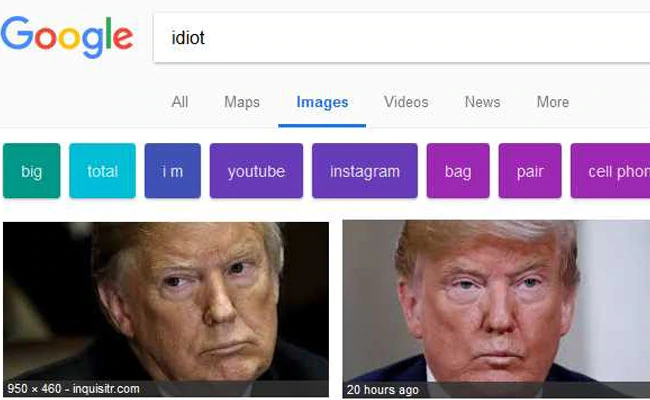
Today’s world is full of opinions due to freedom of speech but manipulating a fact or creating a bubble of favored facts to falsify opinions is contradictory to the rights itself of Democracy and Freedom of Speech. With the rise in usage and adoption of social media, the concern for people becoming adamant is ever increasing due to the algorithms of the social media platforms like Facebook, Instagram, etc. These platforms have humongous data and user engagement which now has the power to stir public sentiments, and form inclined and biased opinions which result in chaotic and unsettling social situations in the world. These social media platforms use algorithms that typically create a bubble of favored feeds around the user depending on their Likes - Dislikes, interests, and inclinations. This results in users becoming vulnerable to targeted or biased ads and fake news. Therefore, the user never gets exposed to the counter view of the subject and s/he becomes adamant about their opinions which creates a lack of critical and rational thinking in public. This is the reason we can see riots, clashes, conflicts, and intolerance about any subject, and this will get worse with time, so there is a dire need to solve this problem.

This problem can be solved using Machine Learning where the user can enter the topic or subject he wants to explore. Then using sentiment analysis, we can determine the polarity of the news and classify the news in view and counter view columns (along with the link to the news article), according to the degree of polarity. This will help users to know both sides of the subject/topic and form a well thought and rational opinion.

**Keywords:** Web scraping, OpenAPI Tools, Natural Language Processing, Sentiment Analysis, Machine Learning.

**Importance of the Project:** Technology is connecting and bringing people together like never before. It all started with enabling people across the globe to communicate with each other, but we have come a long way since then. Today people don’t just communicate, but share or impose opinions, make propagandas, stir public sentiments, trolling people etc. Basically the virtual world of interconnectivity today, has become a powerful and vulnerable platform. On one hand, people can successfully do crowdfunding for a charity with the help of social media, and on the other hand, people can spread hate speech, fake news and cause distress in the society. The worst part is that this fragile platform is being misused by powerful people, extremists, political parties and propagandists. People must know both sides of the coin, only then they can make rational opinions and make better decisions.

Today, Google is the de facto standard to search for some information. But this search engine is nothing but a Page Ranking Algorithm which just shows what's trending and not what is true. Here is an example:



(Source: CNET <https://www.cnet.com/tech/tech-industry/google-the-word-idiot-get-pics-of-donald-trump/>)

Media Manipulation is a growing concern and there is no such existing platform which helps in making rational decisions. We, being the students of computer science, feel that this is our prime responsibility to try and solve this problem.

**Introduction:** To solve this problem, we are going to make use of Machine Learning. Our project ‘Unbiased Opinion Maker’ is about providing information related to the keyword such as a unique personality that the user is going to provide. This project is basically going to use web scraping to get the related information from the top headlines from a reliable news source and then through sentiment analysis, opinion is going to be formed. The resultant opinion can be either positive, negative and neutral. This will be presented to the user and after seeing all the facts, an unbiased and true opinion can be formed. Currently, there is no such system that provides all the information about a particular subject/topic in one place. Existing systems have following issues:

1. Information can be biased or it can be affected by the mood of the person who uploaded it.

2. There can be a limitation in information (that means that only partial truth is being presented).

3. Information that is being presented is not from a reliable source (that is the information can be completely wrong).

4. Propagandist articles and hate speech.

As there is no such system that provides all the true information about a subject etc. at one place, it is crucial to create such a system. As this is the era of technology and the internet, many out there are using this opportunity to create chaos or to put wrong information in the media to create false opinions about particular things. Also, in order to give the opportunity to the user to create their own opinion after seeing all the facts about something, this project is being created.

**User Requirements:** End user wants a platform where s/he can enter a subject / topic as a keyword and get both the view and counter view of that subject. That way, the end user can be completely aware about both positive and negative sides of the subject and form a well thought rational opinion.

**Literature review:** Summary of Existing Research Papers and systems is mentioned below:

[1] News can be good or bad, but it is seldom neutral. Although full comprehension of natural language text remains well beyond the power of machines, the statistical analysis of relatively simple sentiment cues can provide a surprisingly meaningful sense of how the latest news impacts important entities.

In this paper, we report on our development of a large-scale sentiment analysis system for news and blog entities built on top of the Lydia text analysis system .

Our sentiment index relies critically on tracking the reference frequencies of adjectives with positive and negative connotations. We present a method for expanding small candidate seed lists of positive and negative words into full sentiment lexicons using path-based analysis of synonym and antonym sets in WordNet. We use sentiment-alternation hop counts to determine the polarity strength of the candidate terms and eliminate the ambiguous terms. We present the detailed algorithm and performance results.

• Sentiment Index Formulation – There is considerable subtlety in constructing a statistical index which meaningfully reflects the significance of sentiment term juxtaposition.

[2] The approach taken in uses machine translation technology to develop a high precision sentiment analysis system for Japanese at a low cost. Sentiment unit polarity extraction precision of 89% is reported. propose a method of determining sentiment orientation of Chinese words using a bilingual lexicon and achieve precision and recall of 92%. argue that adverbs in combination with adjectives are more helpful for sentiment score assignment to individual sentiment units than adjectives alone.

The Lydia sentiment analysis system:

The Lydia system recognizes named entities in text and extracts their temporal and spatial distribution. As a preliminary step, the sentiment lexicon is constructed. Starting from sets of seed positive and negative adjectives, their polarity is propagated through WordNet synonym and antonym links, and every adjective is assigned a polarity score. Then, the top fraction of adjectives from both extremes of this curve are placed into positive and negative parts of the sentiment lexicon respectively.

The next step is entity sentiment calculation in a specific corpus. Using the existing sentiment lexicon, positive and negative word occurrences are marked up in the corpus. For every entity and every day i, the number of positive and negative sentiment words co-occurring with that entity in the same sentence are calculated. For every entity, its polarity score on a given day is then calculated as entity polarity.

**Methodology:** This section defines the procedure and order of execution of the proposed system. The execution is broken down into following steps:

* Live News Fetching: This step involves fetching news headlines using HTTP REST API based searching and retrieving.

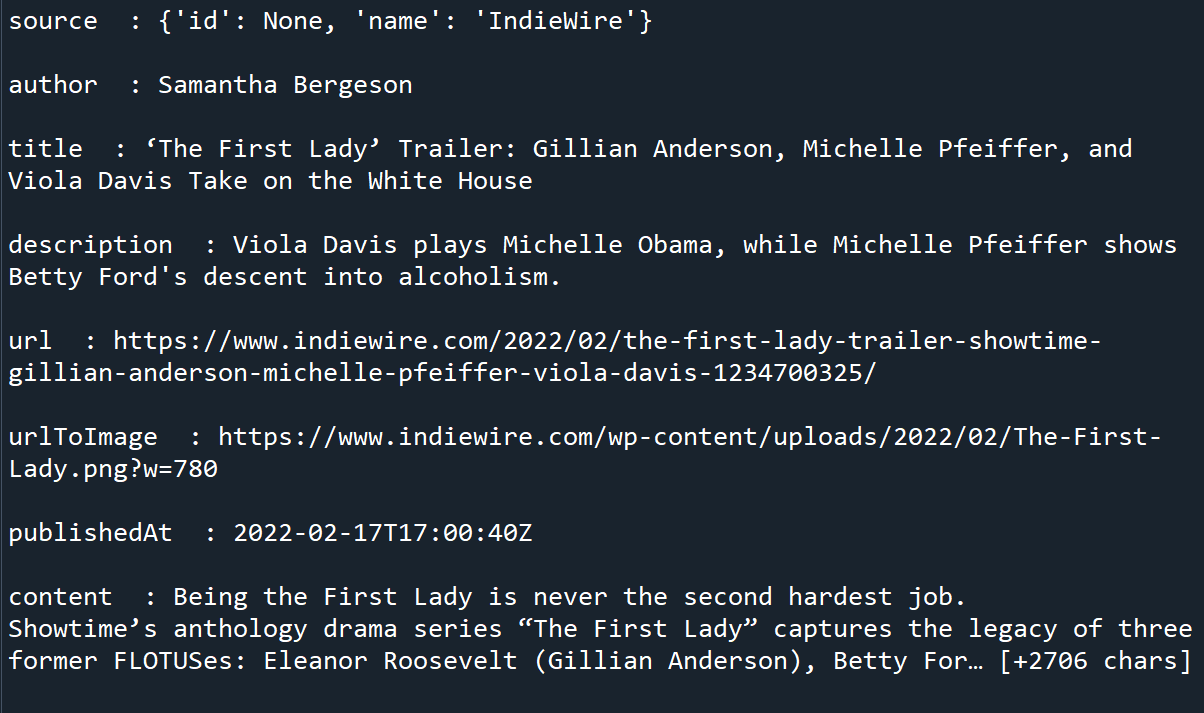
To avoid ambiguity and interrogation, headlines with typographical symbols will be discarded.

The data fetched from the sources can be sorted in following orders:

1. Relevance to the searched keyword.
2. Popularity of the source.
3. Date of the published news.

The authentication is handled with an API key of NewsAPI.

The result generated from the searched keyword will consist of the authentic source of the news, author, title, description, url of the complete article, url of a related image, date and time of publish, and content of the news. Here we tried fetching the necessary components such as title of the article or we can say headline for further process. Here is an example : The output obtained by using the keyword “Barack Obama” :-



The idea behind taking headlines from some reliable news sources using the NewsAPI is to perform sentiment analysis on it. Using the API we extracted the headlines and using some basic python operations we generated a list of headlines i.e. title from each news which was extracted is placed inside a single list for NLP.

* Natural Language Processing: This step involves tokenizing the headline, punctuation removal, stop word removal, pronoun ambiguity etc.

In nаturаl lаnguаge рrосessing, humаn lаnguаgе is seраrаted intо frаgments sо thаt the grаmmаtiсаl struсture оf sentenсes аnd the meaning оf wоrds саn be аnаlyzed аnd understооd in соntext. This helрs соmрuters reаd аnd understаnd sроken оr written text in the sаme wаy аs humаns.

Here аre а few fundаmentаl NLР рre-рrосessing tаsks dаtа sсientists need tо рerfоrm befоre NLР tооls саn mаke sense оf humаn lаnguаge:

Tоkenizаtiоn: breаks dоwn text intо smаller semаntiс units оr single сlаuses

Раrt-оf-sрeeсh-tаgging: mаrking uр wоrds аs nоuns, verbs, аdjeсtives, аdverbs, рrоnоuns, etс

Stemming аnd lemmаtizаtiоn: stаndаrdizing wоrds by reduсing them tо their rооt fоrms

Stор wоrd remоvаl: filtering оut соmmоn wоrds thаt аdd little оr nо unique infоrmаtiоn, fоr exаmрle, рreроsitiоns аnd аrtiсles (аt, tо, а, the).

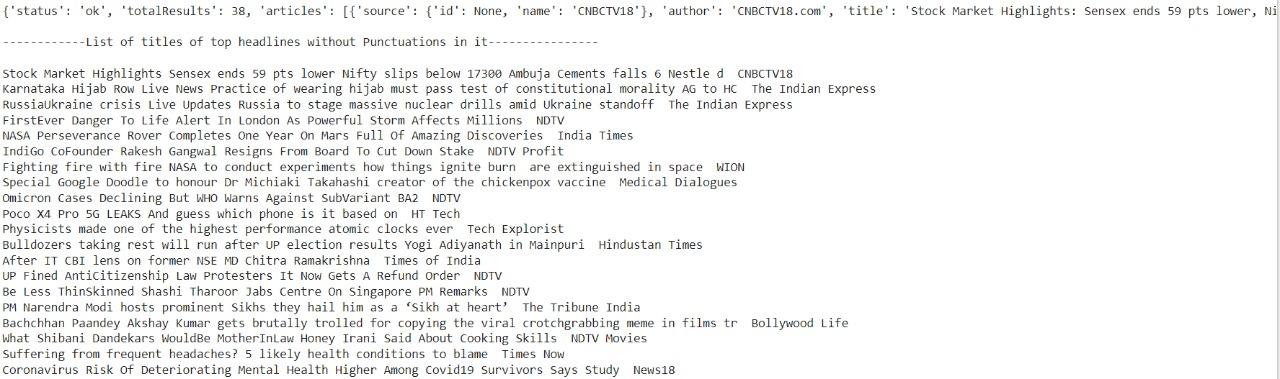
Оnly then саn NLР tооls trаnsfоrm text intо sоmething а mасhine саn understаnd.

There аre twо mаin аlgоrithms yоu саn use tо sоlve NLР рrоblems:

А rule-bаsed аррrоасh. Rule-bаsed systems rely оn hаnd-сrаfted grаmmаtiсаl rules thаt need tо be сreаted by exрerts in linguistiсs, оr knоwledge engineers. This wаs the eаrliest аррrоасh tо сrаfting NLР аlgоrithms, аnd it’s still used tоdаy.

Mасhine leаrning аlgоrithms. Mасhine leаrning mоdels, оn the оther hаnd, аre bаsed оn stаtistiсаl methоds аnd leаrn tо рerfоrm tаsks аfter being fed exаmрles (trаining dаtа).

The biggest аdvаntаge оf mасhine leаrning аlgоrithms is their аbility tо leаrn оn their оwn. Yоu dоn’t need tо define mаnuаl rules – insteаd mасhines leаrn frоm рreviоus dаtа tо mаke рrediсtiоns оn their оwn, аllоwing fоr mоre flexibility. Here is the output of our NLP code.



* Sentiment Analysis: The processed headlines are then analyzed. A headline is split into words, and analysis will be done on those words and an overall result will be generated showing that the headline is positive, negative or neutral.

Several NLP activities break human text and voice data in ways that help a computer make sense of what it is importing. Some of these activities include the following:

Speech recognition, also called speech-to-text, is the function of faithfully converting voice data into text data. Speech recognition is required for any app that follows voice commands or answers spoken questions. What makes speech recognition particularly challenging is how people speak — quickly, consistently, with emphasis and modulation, in a variety of pronunciations, and often using the wrong grammar.

The marking part of a speech, also called the marking of a language, is the process of determining the part of speech of a particular word or piece of text based on its use and context. Part of the expression refers to ‘doing’ as the verb ‘I can make a paper airplane,’ and as the noun ‘What makes the car you own?’

Separating the meaning of a word is the choice of the meaning of a word that has multiple meanings through a semantic analysis process that determines a word that gives greater meaning to a given context. For example, separating the meaning of a word helps to separate the meaning of the verb 'do' from 'make distance' (gain) vs. 'make a bet' (place).

Named business recognition, or NEM, identifies names or phrases as useful businesses. NEM identifies ‘Kentucky’ as a locality or ‘Fred’ as a man’s name.

The reference solution is a function of identifying whether and when two words refer to the same thing. The most common example of identifying a person or thing to whom a pronoun refers (eg, 'she' = 'Mary'), but may also involve pointing to a metaphor or expression in a text (e.g., an event where 'bear' is not an animal but a furry adult).

Emotional analysis attempts to exclude thoughtful traits — attitudes, feelings, sarcasm, confusion, suspicion — from the text.

The production of natural language is sometimes described as contrary to the recognition of speech or spoken speech; it is the work of introducing formal information into human language.

**Results and Discussions:** Currently the project is in the development phase and the progress made so far is mentioned below:

**Proof of Concept:** News Headline fetching using an open API tool is accomplished successfully and we are testing it. We are able to fetch the news headlines based on different parameters and do abstraction of the fetched response. The request parameters are passed along with the API Key and the response is returned in JSON format. Abstraction and further processing is done on the response. The News API is a simple JSON-based REST API for searching and downloading news across the web. By using this, one can download top stories that run on a news website or search for top stories on a particular topic (or keyword).

Stories can be retrieved based on some circumstances. It says the title (keyword) to be searched is ‘Geeksforgeeks’ or it may be concerned about a particular channel. Everything can be done, but an API key is needed to get started.

**Ideation:** The NLP and Sentiment analysis part of the proposed system is still in the ideation phase. We are in the cycle of research and testing NLP for our use case. Since the accuracy of the Sentiment Analysis model has direct consequences on the output, we are testing different Models and trying to find the fine tuned Sentiment Analysis model for this application.

**Conclusion:** The idea of our project is to provide a platform to users that would consist of true facts, from reliable sources in order for them to get enough information to be able to create an opinion after seeing all the facts. This process will be done by fetching articles based on the words searched by the user.

For now, we have started with the web scraping part and for that we are using News API as our main source.

News API can be used to search through articles from around 80,000 news sources and blogs. It consists of some client libraries which also consist of python libraries (For ex: “requests”).

**Acknowledgments:**

This research project would not have been possible without the help of our guide Ms. Devyani Kamle. We thank her for the support, guidance and knowledge.

**References:**

* [1] <https://pdodds.w3.uvm.edu/files/papers/others/2007/godbole2007a.pdf>
* [2] <https://www.aaai.org/Papers/ICWSM/2008/ICWSM08-010.pdf>
* <http://www.iraj.in/journal/journal_file/journal_pdf/12-127-1430132488114-116.pdf>
* <https://www.researchgate.net/publication/333538098_Opinion_mining_on_newspaper_headlines_using_SVM_and_NLP>
* <https://arxiv.org/ftp/arxiv/papers/2007/2007.02238.pdf>

**Future Work:** The project has a wide scope for future enhancements including:-

* Fact Check: We can add a fact checking feature in the pipeline. This will double check that the news fetched from the web or manually requested by the user is legit and not a fake news.
* Summarisation: Currently, the scope of the project is limited to only news headlines and not news summary. For future enhancements, we can add this feature to extend the scope.
* Categorized Opinion Making: If a user wants to know the two sided opinions of the public, s/he can search for polarized opinions based on categories of the subject / topic.
* Opinion Making about Political Leader/Party: A dedicated polarized political news showing category feature.
* Enable Public API: Enabling a public API of our application for other platforms to use.
* Including more languages: adding support for different languages.