**Positive and Negative Sentiments**

**A screenshot of a cell phone

Description automatically generated**

1. Above plot shows number of posts with -ve sentiment against the subreddits.
2. The graph formed is very steep which means people don’t reply to negative posts which results in **low engagement**.
3. By reply we mean that consider a scenario where Subreddit A writes negative posts targeting Subreddit B and Subreddit B **did not reply**. Meaning A is the source Subreddit and B is target Subreddit, when it comes to reply B should source and A the target.
4. This scenario is more with -ve sentiments that’s why the slope of the graph is less.
5. Also, highest no. of -ve posts are above 1400.

A close up of a mans face

Description automatically generated

1. Above plot shows number of posts with +ve sentiment against Subreddits.
2. The graph formed is not that steep when compared with above graph which means people reply to positive posts resulting in **high engagement**.
3. By reply we mean that consider a scenario where Subreddit A writes positive posts targeting Subreddit B and Subreddit B **did reply**. Meaning A is the source Subreddit and B is target Subreddit, when it comes to reply B should source and A the target.
4. This scenario is more with +ve sentiments that’s why the slope of the graph is more compared to the plot with -ve sentiments.
5. Also, highest number of +ve posts is above 3000.

**Likely Number of males and females in Subreddits**

A close up of a map

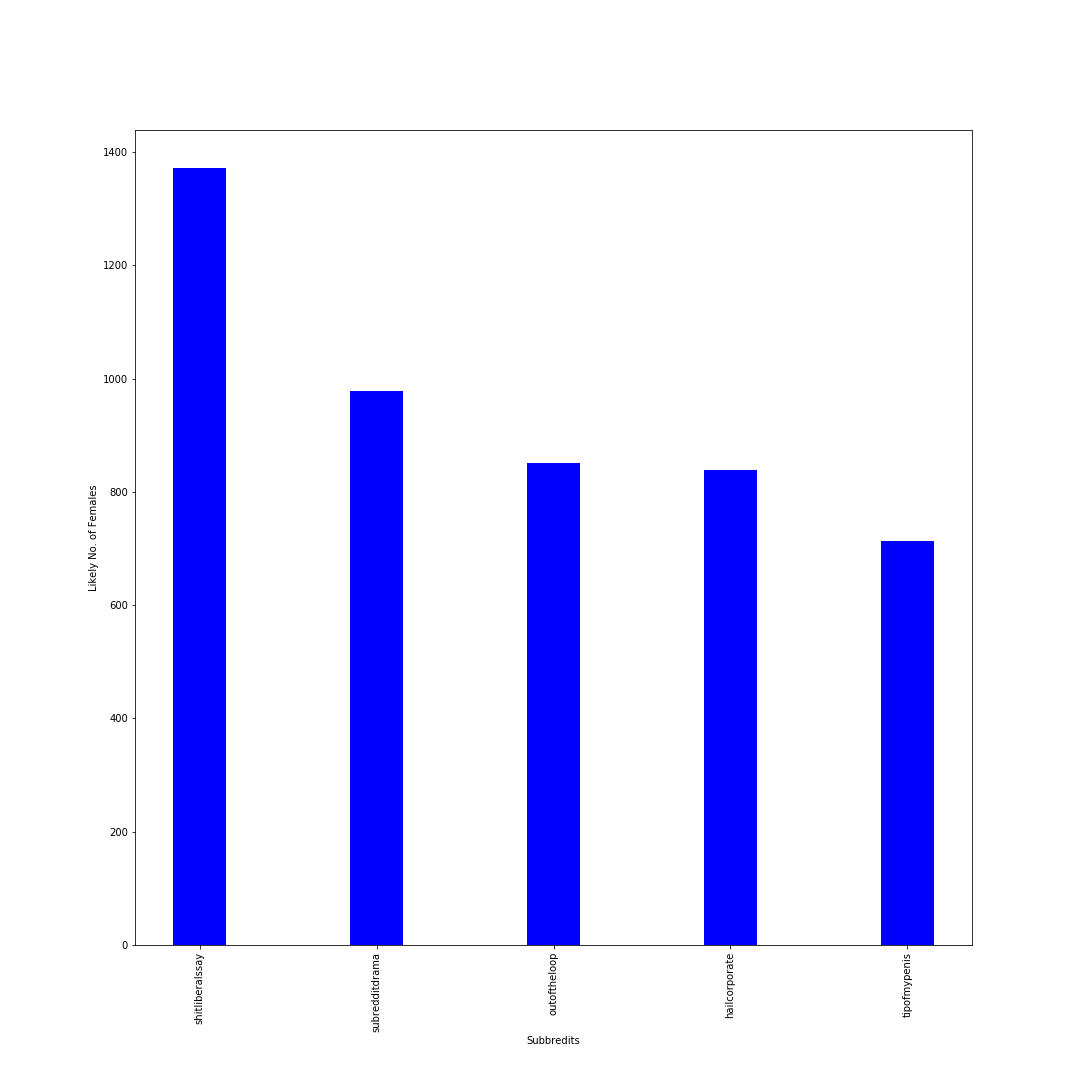
Description automatically generated

1. Based on LIWC TexProps we tried to classify that whether a post was written by a male or a female user. The LIWC properties that we considered for **females were 23 27 28 29 and males were 38 31 25 26**.
2. Females tend to use more pronouns, 2nd and 3rd person narrative whereas males use more articles, prepositions and 1st person narrative in their speech or while writing.
3. Males and females have different speech styles as talked by **Yla R. Tausczik and James W. Pennebaker** in **The Psychological Meaning of Words: LIWC and Computerized Text** [1].
4. Additional documents “Male and female speech styles.pdf” and “LIWC2007LanguageManual.pdf” we will be attaching along with the assignment for more reference.
5. From this plot we can see that number of males and females have increased over the years and is still growing. Scaling the data and including data for years 2017,2018 and 2019 would also show similar trends.

A close up of a logo

Description automatically generated

1. Above pie chart shows that reddit has **81.2% male users and 18.8% female** users according to the dataset provided which is quite close to the actual numbers which are **71% males and 29% females**.[2].
2. Scaling the data and including more Subreddits and data from more years, its expected the we will get **similar results**.
3. Also, it is our attempt to **create a very basic heuristic model** which classifies whether the user is a male or a female based on TextProps provided [1].



1. Above graph shows top 5 subreddits with highest number of females.
2. Shitliberalssay has the highest number of female users according to our prediction.
3. Highest number of females is around 1300 according to our model.

A close up of a logo

Description automatically generated

1. This graph shows top 5 subreddits with highest number of males.
2. Subredditdrama has the highest number of males.
3. Highest number of males is above 3500 according to our model.

**If the Post is Deceptive or Not**

**A screenshot of text

Description automatically generated**

1. Based on LIWC TexProps we tried to classify that whether the post is deceptive or honest. By deceptive we mean that if the post is misleading. The LIWC properties that we considered to classify the post as **Deceptive are 73 50 62 25.**
2. Deceptive texts have more motion words, negative emotions, exclusion words and use 1st person narrative.
3. Deceptive texts have some properties as talked by **Yla R. Tausczik and James W. Pennebaker** in **The Psychological Meaning of Words: LIWC and Computerized Text** [1].
4. Additional documents “LIWC2007LanguageManual.pdf” we will be attaching along with the assignment for more reference.
5. From this plot we can see that number deceptive/misleading posts have increased over the years and is still growing. Misleading information is a major problem of today’s information Era and models and algorithm are being designed to recognize them, for example Facebook implemented AI Algorithm to automatically detect and delete Fake news and misleading information.

A close up of a logo

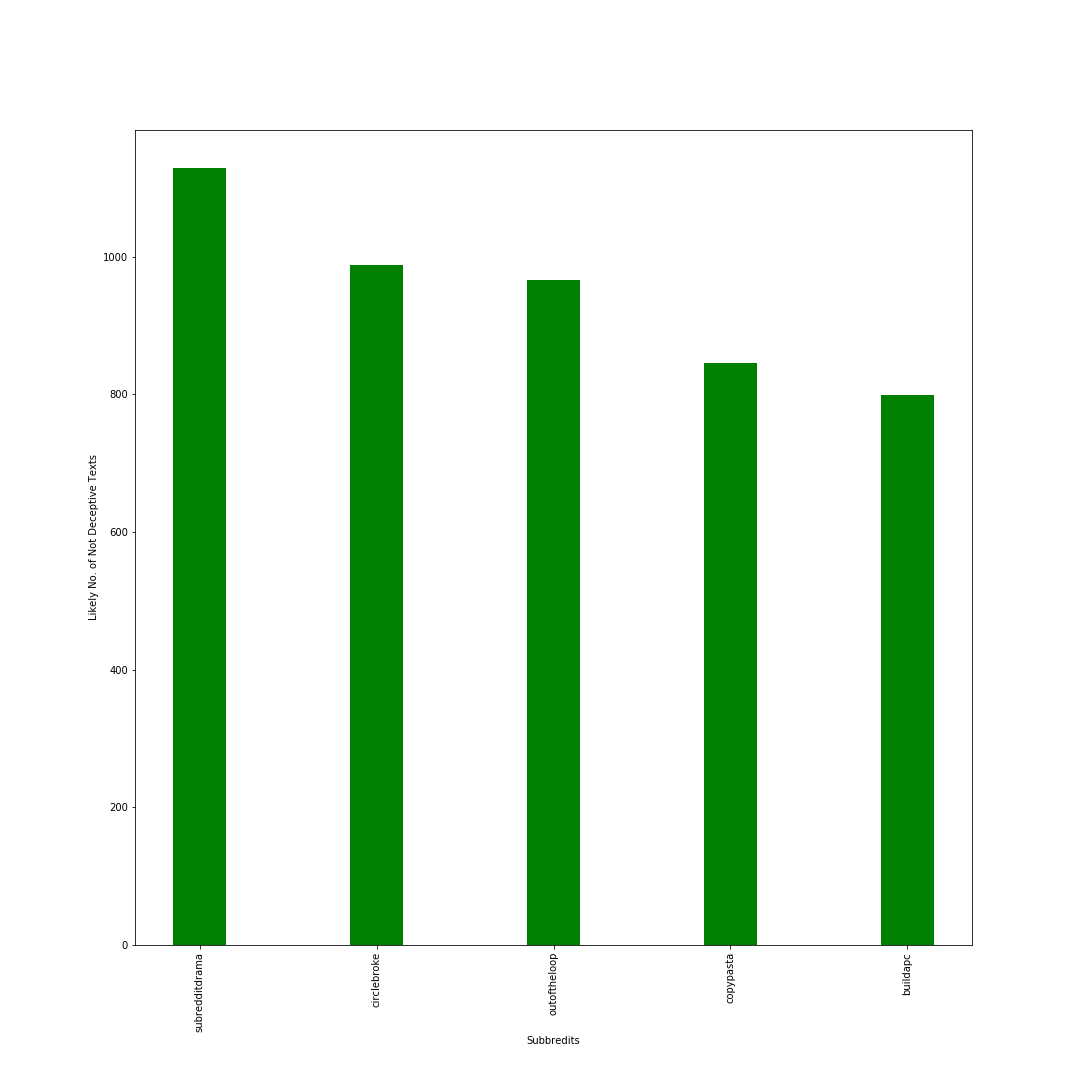
Description automatically generated

1. Above plot depicts that 37.1% of the posts are deceptive/misleading and 62.9% are honest.
2. It is our attempt to **create a very basic heuristic model** which classifies whether the post is deceptive or not based on TextProps provided [1].
3. Such advanced models are necessary to filter out fake news and misleading posts so that social media and other websites are a safe and informative experience.
4. Fake news has been the cause of lot problems all around the world and its mostly circulated through social media.

A close up of a logo

Description automatically generated

1. This plot shows that subredditdrama has most deceptive posts. This subreddit is about people writing posts about what’s going on in other subreddits. Subredditdrama has most -ve sentiment posts because people there tend enjoy the fights going around, so it’s natural for this subreddit to have highest number of deceptive posts.
2. Also, conspiracy subreddit is at 4th number which seems fair because in this subreddit people post conspiracy theories which can be untrue and misleading.



1. Subbreditdrama has the highest number of non-deceptive posts but these are less the deceptive posts as subredditdrama has the most deceptive posts too.
2. Buildapc subreddit is about to building custom computers which are more facts that’s why Buildapc has high non-deceptive posts.

Research Paper

A) We have taken references from **The Psychological Meaning of Words: LIWC and Computerized Text** by **Yla R. Tausczik and James W. Pennebaker.**

We can tell a lot of things from a person’s speech and writing style, for example maturity level, sex, psychological processes, status etc.

We have done few visualizations using this information.

Whether the person posting is male or female, also whether the post is deceptive or not.

These visualizations take reference from the research paper mentioned above. The author clearly explains how the above information from a person’s text or speech can help us know a lot of things, such techniques are also used in lie-detection test too.

B) Also we have used Indexing in this assignment to reduce the time taken for the data to load in Neo4j and create graph from it. The research paper “**Nanosecond Indexing of Graph Data With Hash Maps and VLists**” talks about indexing.

In Neo4j we can create index over the properties on any node that has been given a label. Once these indexes are created Neo4j will manage and update them whenever the database is changed.

When Neo4j creates an index, it creates a redundant copy of the data in the database. Therefore, using an index will result in more disk space being utilized, plus **slower writes to the disk**. Which is also talked about in the search paper. Generally, it's a good idea to create an index when you know there's going to be a lot of data on certain nodes. Also, if the queries are taking too long to return, adding an index may help.