



# Social Media Analysis - NLP Using Glove Embeddings

Shreyas D<sup>1</sup>, Dr Sujithra M<sup>2</sup>

<sup>1</sup>Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.

<sup>2</sup>Department of Computing-Data Science, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.

## How to cite this paper:

Shreyas D<sup>1</sup>, Dr Sujithra M<sup>2</sup>, "Social Media Analysis - NLP Using Glove Embeddings", IJIRE-V3I06-118-121.

Copyright © 2022 by author(s) and 5<sup>th</sup> Dimension Research Publication.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>

**Abstract:** As the scourge of "fake news" continues to plague our information environment, attention has turned toward devising automated solutions for detecting problematic online content. But, in order to build reliable algorithms for flagging "fake news," we will need to go beyond broad definitions of the concept and identify distinguishing features that are specific enough for machine learning. With this objective in mind, we conducted an explication of "fake news" that, as a concept, has ballooned to include more than simply false information, with partisans weaponizing it to cast aspersions on the veracity of claims made by those who are politically opposed to them. We identify seven different types of online content under the label of "fake news" (false news, polarized content, satire, misreporting, commentary, persuasive information, and citizen journalism) and contrast them with "real news" by introducing a taxonomy of operational indicators in four domains—message, source, structure, and network—that together can help disambiguate the nature of online news content.

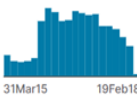
**Key Word:** NLP, Data Visualization, CNN, sentiment classification, document clustering.

## I. INTRODUCTION

Fake news refers to misinformation, disinformation or mal-information which is spread through word of mouth and traditional media and more recently through digital forms of communication such as edited videos, memes, unverified advertisements and social media propagated rumors. "Fake news," or fabricated information that is patently false, has become a major phenomenon in the context of Internet-based media. It has received serious attention in a variety of fields, with scholars investigating the antecedents, characteristics, and consequences of its creation and dissemination. Some are primarily interested in the nature of misinformation contained in false news, so that we can better detect it and distinguish it from real news. Others focus on the susceptibility of users—why we fall for false news and how we can protect ourselves from this vulnerability. Both are geared toward improving media literacy to protect consumers from false information.

## II. DATASET

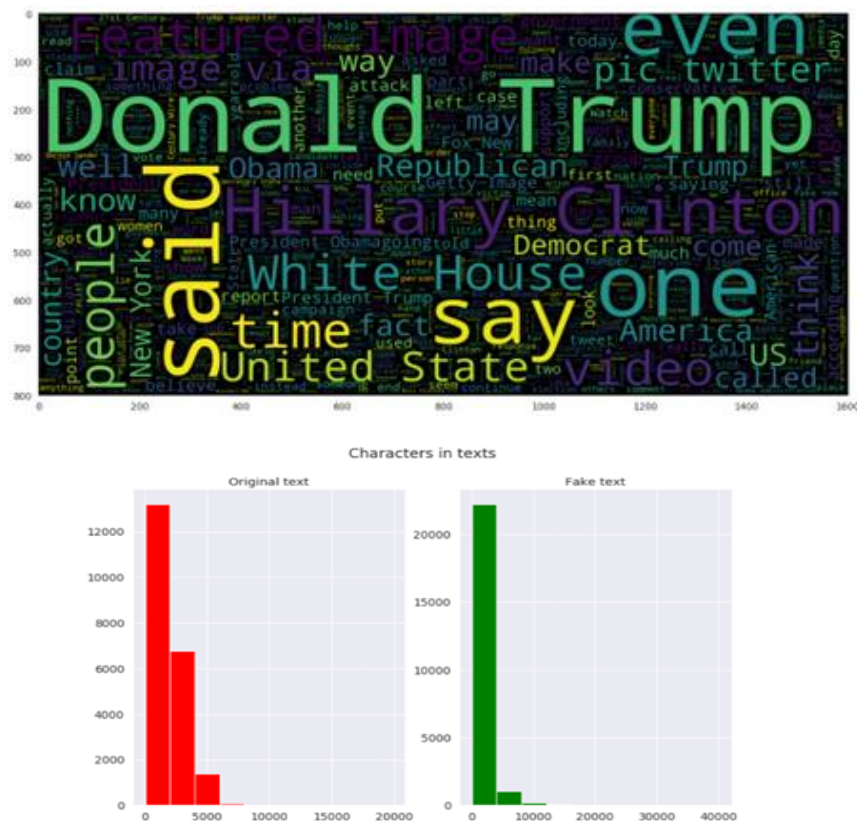
This dataset consists of about 40000 articles consisting of fake as well as real news. Our aim is to train our model so that it can correctly predict whether a given piece of news is real or fake. The fake and real news data is given in two separate datasets with each dataset consisting around 20000 articles each.

A title	A text	A subject	date
The title of the article	The text of the article	The subject of the article	The date at which the article was posted
17903 unique values	[empty] 3% AP News The regul... 0% Other (22851) 97%	News 39% politics 29% Other (7590) 32%	
Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing	Donald Trump just couldn't wish all Americans a Happy New Year and leave it at that. Instead, he had...	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He's been under the as...	News	December 31, 2017

## III. WORD CLOUD

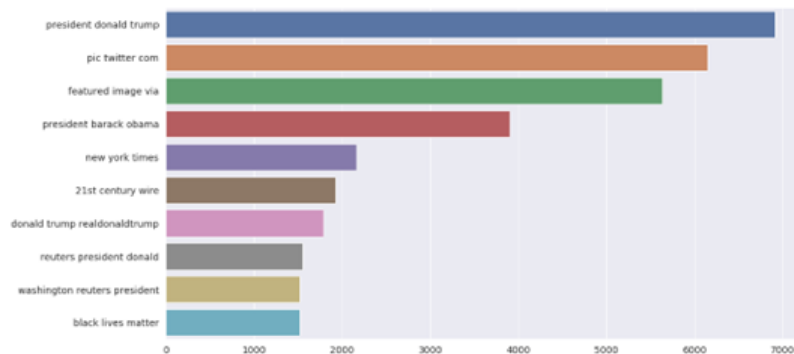
A Word Cloud or Tag Cloud is a visual representation of text data in the form of tags, which are typically single words whose importance is visualized by way of their size and color. As unstructured data in the form of text continues to see unprecedented growth, especially within the field of social media, there is an ever-increasing need to analyze the massive amounts of text generated from these systems. A Word Cloud is an excellent option to help visually interpret text and is useful

in quickly gaining insight into the most prominent items in a given text, by visualizing the word frequency in the text as a weighted list.



#### IV. UNIGRAM ANALYSIS

N-grams are continuous sequences of words or symbols or tokens in a document. In technical terms, they can be defined as the neighboring sequences of items in a document. They come into play when we deal with text data in NLP(Natural Language Processing) tasks.



#### Removing Stop Words and Punctuation

A few tokens are less significant than others. For example, familiar words, "the" probably won't be extremely useful for uncovering the fundamental qualities of a text. So ordinarily it is smart to kill stop words and accentuation marks before doing an advanced investigation.

#### Computing Term Frequency or TF-IDF

After preprocessing the text data, you can then keep on creating features. For document clustering, one of the most broadly perceived approaches to making features for a record is to sort out the term frequencies of all of its tokens. Though defective, these frequencies can by and large give a couple of bits of knowledge about the report's subject.

#### V. GLOVE METHOD

Glove method is built on an important idea: we can derive semantic relationships between words from the co-occurrence matrix. Given a corpus having  $V$  words, the co-occurrence matrix  $X$  will be a  $V \times V$  matrix, where the  $i$  throw and  $j$  th column of  $X$ ,  $X_{ij}$  denotes how many times word  $i$  has co-occurred with word  $j$ . An example co-occurrence matrix might

look as follows.

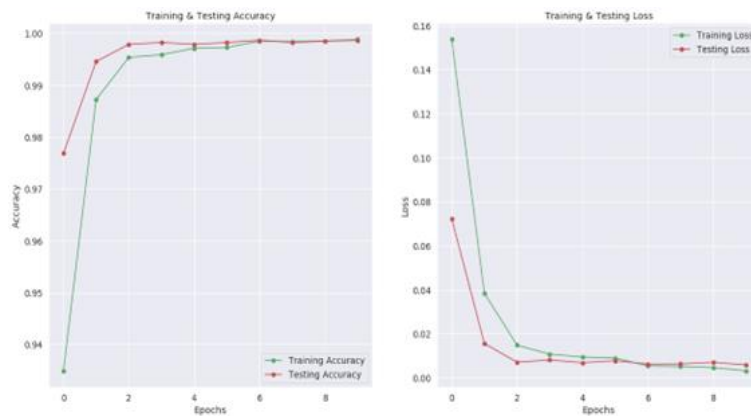
	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

The co-occurrence matrix for the sentence “the cat sat on the mat” with a window size of 1. As we probably noticed it is a symmetric matrix. How do we get a metric that measures semantic similarity between words from this. For that, we will need three words at a time.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k ice)$	$1.9 \times 10^{-4}$	$6.6 \times 10^{-5}$	$3.0 \times 10^{-3}$	$1.7 \times 10^{-5}$
$P(k steam)$	$2.2 \times 10^{-5}$	$7.8 \times 10^{-4}$	$2.2 \times 10^{-3}$	$1.8 \times 10^{-5}$
$P(k ice)/P(k steam)$	8.9	$8.5 \times 10^{-2}$	1.36	0.96

## VI. MODEL PERFORMANCE

After preparing the model, next decided on the text precision utilizing the best loads of the prepared model. The test accuracy acquired was around 88%. It's sufficiently not to have a model that performs well as indicated by a given measurement: you should likewise have a model that you can comprehend and whose results can be made sense of. Begin by preparing the model on a piece of the dataset, and afterwards break down the primary wellsprings of misclassification on the test set.



	precision	recall	f1-score	support
Fake	1.00	1.00	1.00	5858
Not Fake	1.00	1.00	1.00	5367
accuracy			1.00	11225
macro avg	1.00	1.00	1.00	11225
weighted avg	1.00	1.00	1.00	11225

## VII. CONCLUSION

Social Media classification is uncommonly significant in recovering explicit data in a set period. Regardless, a beneficial instrument can reduce the expense and time of searching for and recovering important data. Whereas the model worked more than expected, all credits goes to the glove’s method for finding the fake and real texts in social media context view.

### References

- [1] J Jeremy Howard sylvain gugger. "Deep Learning for Coders with fastai and PyTorch" O'Reilly media, Inc. July 2020
- [2] Johnson Kolluri, Shaik Razia, and Soumya Ranjan Nayak. "Text classification using machine learning and deep learning model", in arxiv.org, June 2020
- [3] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 7. "Deep learning based text classification: A comprehensive review", in arxiv.org, January 2022
- [4] muthana AI-Amidie . Laith ALzubzidi." Review of deep learning : CNN architectures, challenges, application of future directions." in journal of big data 31 march 2021
- [5] Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. "A comparative analysis of logistic regression, random forest and KNN models for the text classification", March 2020
- [6] J Jeremy Howard sylvain gugger. "Deep Learning for Coders with fastai and PyTorch" O'Reilly media, Inc. July 2020
- [7] Johnson Kolluri, Shaik Razia, and Soumya Ranjan Nayak. "Text classification using machine learning and deep learning model", in arxiv.org, June 2020
- [8] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 7. "Deep learning based text classification: A comprehensive review", in arxiv.org, January 2022
- [9] muthana AI-Amidie . Laith ALzubzidi." Review of deep learning : CNN architectures, challenges, application of future directions." in journal of big data 31 march 2021
- [10] Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. "A comparative analysis of logistic regression, random forest and KNN models for the text classification", March 2020