Sentimental Analysis of Twitter tweets and Reddit posts

Ankit Devri
School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab, India
vivekdevri@gmail.com

Abstract- *Microblogging has become an integral part of our daily lives in this digital age. People post their thoughts on platforms such as Twitter and Reddit on the daily basis. In order to estimate the opinion of the people companies nowadays uses Sentiment Analysis technique on the posts of the people on the platform to judge the opinion of people. These opinions help them in forming the strategy for the future. There as many methods and state of the art models proposed by researchers for sentiment analysis. Techniques such as stemming, lemmatization, POS features, lexicon features help in sentiment analysis of a statement. There are many state-of-the- art baselines for sentiment analysis such as the n-grams method, bag of words, word2vec, TFIDF and many more models have been produced. This research paper investigates about the best methods and models in sentiment analysis of Twitter tweets and reddit posts*

Keywords- *n-grams, bag of words, POS, lexicon, TFIDF, stemming, lemmatization, word2vec*

I. Introduction

In the recent decade there has been a shift in how people communicate with each other throughout the world. There is an emergence of digital age where everybody is connected with each other online using many different platforms available on the internet. Earlier it used to happen that we could only communicate by phone. But now due to emergence of internet and people having access to the internet there has been a huge change in way we communicate. Emergence of microblogging services such as Facebook, Reddit and Twitter have made people aware that how they can communicate with the entire community by their device.

During the covid pandemic, there has been a huge growth on usage of these platforms. Microblogging websites have evolved to become a source of varied kind of information. This is due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. During the pandemic time one of the only ways to stay connected was through these microblogging sites. This resulted in generation of great amount of sentiment on a specific topic on the platform. As the demand surged the amount of data exploded, growing larger and larger every day.

This resulted in the microblogging sites such as Twitter and reddit to have huge amount of data related to different things. These companies devised a way to use the data by applying Sentiment Analysis of the provided data on that topic. The analysis thus served as feedback of a person on a topic. Companies use this feedback on any of their services to make decisions for the future. However, in case of Twitter and Reddit there are a lot of varied topics covered and there are a lot of opinions of people on them. The breadth of the data available is enormous in both the platforms. Twitter also has special way of denoting the sentiment of a statement using hashtags which people use whenever posting. Hashtags and emoticons are one of the common ways to express one's opinion on microblogging websites.

This paper mentions different techniques and models used by researchers in sentiment analysis and which model can provide best use case for a sentiment analysis problem. Natural language processing is integral part of solving real life problems. In our experiments we used a dataset extracted from Kaggle and performed sentiment analysis on it. The data was cleaned thoroughly and tested on three major models: Linear SVC model,
Naïve Bayes and Ensemble model. Supervised Learning techniques were used along with polarity and subjectivity with the ensemble model.

There are many features present in a twitter tweet which can help determine the opinion. POS tagging, Polarity, Hashtags, Emoticons, abbreviations, Lexicon features are some of them. Looking at different techniques and methods it will be determined which features bring in the maximum effect and which features should be used for which use case.
The paper aims to identify which features to use for a specific use case. These techniques will be compared with the state-of-the-art models proposed earlier

like the n-gram baseline model, naïve bayes base line models.

II. Literature Review

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and the phrase level (Wilson et al., 2005; Agarwal et al.,2009). Microblog like Twitter and Reddit on which users post their real time reactions on varied topics poses newer challenges which appear to be difficult. There are varied types of text data some contains humour, sarcasm and other Parts of Speech features which are hard to detect even for a well-trained machine learning model.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau et al, 2011 [1] used POS specific prior polarity features in their tree kernel to outperform the n-gram baseline. A unigram model is indeed a hard baseline achieving over 20% over the chance baseline for both classification tasks. Their feature base model that uses only 100 features achieves similar accuracy as the unigram model that uses over 10,000 features. The tree kernel-based model outperformed unigram model by a significant margin. It used the POS features and defined specific polarity to each of the emoticons and used English based translations of acronyms typically used by people in typical posts. The unigram model achieved accuracy of about 75% which acted as a baseline and the tree kernel outperformed the unigram and the Semi-features by 2.58% and 2.66% respectively. The 100 Semi-features described in performs as well as the unigram model that uses about 10,000 features. They also experimented with combination of models. Combining unigrams with Semi-features outperforms the combination of kernels with Semi-features by 0.78%.

Efthymios Kouloumpis, Theresa Wilson, Johanna Moore et al, 2011 [2] used different datasets to experiment with their work. They used hash tagged dataset and emoticons dataset to determine usefulness of the features. They also used words listed the MPQA subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2009) and marked them with a prior polarity. There experiment was to study the usefulness of the features such as the n-gram features, hashtags, emotes, parts of speech features, micro blogging features and lexicon features for sentiment analysis. They used n-gram model with combination of all these different features to see the difference in level of performance. They saw those parts of speech features were not that useful in their experiment, while compared to that the lexicon features did have an impact on the performance of the model. However, the most useful of them all was the microblogging features such as hashtags and emotes which made significant impact on the performance. Though it depends on the input data whether it has these features in them itself to extract proper sentiment through them. Hashtags and emotes are applicable for twitter, while only emotes is applicable for Reddit.

Alec Go, Lei Huang, Richa Bhayani et al, 2009 [3] used a Naïve Bayes Classifier built from scratch and imported maximum entropy model, support vector machine from other libraries. They trained the model using unigrams and bigram features, also using feature selection process using frequency-based feature selection, mutual information and x^2 feature selection with the unigram features. The Naïve Bayes classifier outperformed the maximum entropy model and SVM. They also extended the Naïve Bayes with subjective and objective classes which failed due to addition of noisy data. It gave them about 82% accurate results. They took note of various outliers and mentioned them in their paper which showed much work is needed to make a custom model for each natural language task as it came out be very domain specific. Their SVM model could only achieve 73.913% accuracy. The use of POS taggers and other lexicon features is recommended to further increase the accuracy. There model saw accuracy increase of 6%. There was a special difficulty in their model to identify a neutral statement as it contained both positive and negative sentiment in it. Their model could not handle statement with not as it reversed the polarity of the sentiment. It was concluded by them to go for ensemble methods for sentiment analysis.

Hassan Saif, Yulan He, and Harith Alani et al, 2012 [4] proposed to extract sentiment concepts from the datasets they were working on which were Stanford Twitter Sentiment Corpus (STS), Health Care Reform (HCR), Obama-McCain Debate (OMD).Measuring correlation of semantic concepts extracted from the datasets with negative/positive sentiment which then were incorporated to sentiment classification for their model. They used Naïve Bayes Classifier as the base model to incorporate the extracted sentiment concepts on. They proposed three different methods for incorporation.Semantic Replacement is a method in which they replace all entities in tweets with their corresponding semantic concepts. This leads to the reduction of the vocabulary size. Semantic Augmentation is a method augments the original feature space with the semantic concepts as additional features for the classifier training. Semantic Interpolation which is more principal way to incorporate semantic concepts is through interpolation where they interpolated the unigram language model in NB with the generative model of words given semantic concepts. It was observed that semantic interpolation gave the most accurate results

outperforming the other two method by 6% to 10% gain in accuracy. For each of the datasets used they observed semantic interpolation with Naïve Bayes outperformed for all unigrams baselines and POS baselines. Semantic interpolation produced around 80% of accuracy as the average of all the datasets used.

Alexander Pak, Patrick Paroubek et al, 2010 [5] collected the tweets from the twitter api as their dataset. They used Parts of speech tagging technique to separate subjective and objective data and balanced the dataset to maintain the homogeneity of the dataset. In the feature extraction part, they tokenized the sentence, remove stopwords and constructed n-gram features from the dataset. They used SVM and Naïve Bayes for it and found Naïve Bayes outperformed SVM. They trained two NB one with n-gram features and the other with POS features. They used Sharon Entropy to remove commonly appearing n-grams in the dataset and put only n-grams with low entropy to determine a sentiment. In the second method they used a term salience which is a measure takes a value between 0 and l. The low value indicates a low salience of the n-gram, and such an n-gram should be discriminated. Same as with the entropy, they controlled the performance of the system by tuning the threshold value theta. In between the two methods the salience method outperformed the entropy method.

David Zimbra, Ahmed Abbasi, Daniel Zeng and Hsinchun Chen et al, 2018 [6] is a research paper which aims at benchmarking the results of Sentiment of Twitter data and finding best methodologies for sentiment analysis at different sectors or domains. As mentioned earlier, Twitter is a microblogging website that people of varied background and knowledge use. This creates a wide range of topics to explore and do sentiment analysis on which makes it difficult to achieve accuracy in all the tweets of twitter by training a single model. They discussed different methodologies and features related to different topics and benchmarked these to find most useful and effective of the mentioned. In their thorough research they came across some important conclusions. Importance of Using/Developing Systems that Support Domain Adaptation helped a lot of models achieve more accuracy. The top four performing models used ensemble methods which shows effectiveness of Ensemble Methods. Importance of Including an Array of Lexicons and Linguistic Resources came out as one of the key factors in improving performance from a general model to domain specific model which achieved accuracy up to 90%. Yet they discovered several challenges in twitter sentiment analysis. A General model at best could only achieve 71% accuracy.

Abdullah Alsaeedi, Mohammad Zubair Khan et al, 2019 [7] discussed diverse techniques for Twitter sentiment analysis methods were, including machine learning, ensemble approaches and dictionary (lexicon) based approaches mentioned by different researches int both supervised and unsupervised domain. In addition, hybrid and ensemble Twitter sentiment analysis techniques were explored. Research outcomes demonstrated that machine learning techniques; for example, the Support Vector Machines and Multinomial Naïve Bays Classifier produced the greatest precision, especially when multiple features are included. SVM classifiers may be viewed as standard learning strategies, while dictionary (lexicon) based techniques are extremely viable at times, requiring little efforts in the human-marked archive which are now readily available in sentiment analysis of twitter topic. Machine learning algorithms, such as The Naïve Bayes, Maximum Entropy, and SVM, achieved an accuracy of approximately 80% when n-gram and bigram model were utilized. Ensemble and hybrid-based Twitter sentiment analysis algorithms tended to perform better than supervised machine learning techniques, as they were able to achieve a classification accuracy of approximately 85%.

Lei Zliang, Riddhiman Ghoslr, Mohamed Dekhil, Meichun Hsu, Bing Liu et al, 2011 [8] used the lexicon feature approach for sentiment analysis. They used SVM as their classifying algorithm. They utilized the microblogging features such as the emoticons and the hashtags using binary features instead of frequency, with unigram features also present with them. They conducted an empirical evaluation with the following models. ME: a state-of-the-art learning-based method used by the website “Twitter Sentiment” which uses Maximum Entropy as the supervised learning algorithm. FBS: a lexicon-based method proposed in (Ding et al, 2008) for feature-based sentiment analysis. AFBS: the augmented lexicon-based method for tweets without utilizing the final SVM sentiment classifier. LLS: After Opinion indicators are identified, they put them into the original general opinion lexicon, and run AFBS again. LMS: the model they proposed.
The experiment was run on 5 different datasets and on average LMS came with average accuracy of around 85.4% which was 5% better than LLS, 6% better than AFBS and FBS while being 9% better than ME, thus outperforming every model with a significant margin.

III. Comparative Study

In this section, a comparative study about all the mentioned methods and models has been displayed on Table below.

Comparison of different methods discussed in the literature review

REFERENCE PAPER	DATASET	MODEL	ACCURACY
[1]	Manually Annotated Twitter data	Tree kernel base model + POS + polarity	79%
[2]	Hash tagged + Emoticons + Isieve dataset	n-grams + hashtag + lexicons + emotes	75%
[3]	Manually Annotated Twitter data	Naïve Bayes + lexicon + POS	82%
[4]	Twitter Sentiment Corpus (STS), Health Care Reform (HCR), Obama-McCain Debate (OMD).	Naïve Bayes + Sentiment Replacement, Augmentation, Interpolation	70% 74% 80%
[5]	Manually Annotated Twitter data	Naïve Bayes Entropy, Salience	71%, 78%
[6]	Twitter Api Fetch Manually Annotated data	Ensemble Hybrid (domain specific), Naïve Bayes, SVM (General Purpose)	90% 71%
[7]	Twitter Api Fetch Manually Annotated data	Naïve bayes & SVM, Ensemble Hybrid	80% 85%
[8]	Obama, Harry Potter, iPad, Tangled, Packers	ME, FBS, AFBS, LLS, LMS	74% 77% 79% 80% 85%

IV. Outcome

Most of the researchers have used n-gram as a baseline model to perform model evaluation. Classifiers like SVM, Naïve Bayes were commonly occurring in almost every study. In addition, with Parts of speech features, Hash tagging features, Emoticon Features and Lexicon Features had a very significant role in improving each model's accuracy. Feature Extraction and Selection methods such as the Sentiment Replacement, Sentiment Augmentation, Sentiment Interpolation also impacted the accuracy of supervised machine learning models. Also including the method such as maximum Entropy and Salience contributed in producing more accurate results. Ensemble models were proved to be the best using all the above features in them, they took the first place in achieving the best result. Ensemble Hybrid models in case of domain specific sentiment analysis came out to be extremely accurate while in case of general sentiment analysis it seems more work and research is needed to achieve even greater results.

V. Natural Language Processing Overview

Natural language is what we say and hear every day, it's easy for us to understand natural language but for a machine its very hard to identify even what we are saying. Sentiment Analysis on natural language provides us more automation and advancement towards true AI which can help people in their daily lives. For that to happen machines must understand natural language for better communication.

For that machine learning techniques are used on natural language. Since computer cannot understand words directly the language is processed and standardised for a machine to understand the natural language. This standardization of data must be accurate enough to help the machine train and perform better. In case of sentiment analysis, we saw Naïve Bayes appearing in almost every literature review.

Naïve Bayes Classifier

Naive Bayes is a simple model for classification. It is simple and works well for dividing the text into various categories. We saw multinomial Naive Bayes in most reference papers. It assumes each feature is conditional independent to other features given the class. That is, $P(c \mid t)$. $P(t) = P(c)$. $P(t \mid c)$ where c is a specific class and t is text, we want to classify. P(c) and P(t) is the prior probabilities of this class and this text.
And $P(t \mid c)$ is the probability the text appears given this class. In our case, the value of class c might be POSITIVE or NEGATIVE, and t is just a sentence.
The goal is choosing value of c to maximize $P(c \mid t)$:
where : $P(w_i \mid c)$ is the probability of the i^{th} feature in text t appears given class c. We need to train parameters of P(c) and $P(w_i \mid c)$. It is simple for getting these parameters in Naïve Bayes model. They are just maximum likelihood estimation (MLE) of each one. When making prediction to a new sentence t, we calculate the log likelihood $\log P(c) + \Sigma_i \log P(w_i \mid c)$ of different classes, and take the class with highest log likelihood as prediction.

VI. Future Scope and Conclusion

Sentiment Analysis in AI has a huge scope it future, its use cases and applications are present everywhere there is someone who uses natural language, which are us humans. It's a very first step towards a proper AI. To understand a sentiment in a statement is the first step towards comprehension of natural language. Machine Learning and Deep Learning neural networks are actively playing huge role in making this a reality. Deep learning models like LSTM and Hybrid models can be used in Sentiment Analysis which is what should be worked on going forward.

REFRENCES

[1] Sentiment Analysis of Twitter Data by Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau. Proceedings of the Workshop on Language in social media (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011. c 2011 Association for Computational Linguistics

[2] Twitter Sentiment Analysis: The Good the Bad and the OMG! By Efthymios Kouloumpis, Theresa Wilson, Johanna Moore. Proceedings of the Fifth International AAAI Conference on Weblogs and social media.

[3] Twitter Sentiment Analysis by Alec Go, Lei Huang, Richa Bhayani. Stanford University

[4] Semantic Sentiment Analysis of Twitter Hassan Saif, Yulan He, and Harith Alani Knowledge Media Institute, The Open University, United Kingdom. P. Cudré-Mauroux et al. (Eds): lSWC 2012, Part I, LNCS 7649. PP. 508-524, 2012. © Springer-Verlag Berlin Heidelberg 2012

[5] Twitter as a Corpus for Sentiment Analysis and Opinion Mining Alexander Pak, Patrick Paroubek Université de Pan's-Sud, Laboratoire LlMSl-CNRS, Bétiment 508, F-91405 Orsay Cedex, France

[6] The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation by David Zimbra, Ahmed Abbasi, Daniel Zeng and Hsinchun Chen et al, 2018. ACM Transactions on Management Information Systems, Vol. 9, No. 2, Article 5. Publication date: August 2018.

[7] A Study on Sentiment Analy51s Techniques of Twitter Data by Abdullah Alsaeedi, Mohammad Zubair Khan, Department of Computer Science, College of Computer Science and Engineering Taibah University, Madinah, KSA. (IJA CSA) International Journal of Advanced Computer Science and Applications, Vol. 10, N0. 2, 2019.

[8] Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis by Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu, HP Laboratories, HPL-2011-89. © Copyright 201 1 Hewlett-Packard Development Company, LP.