A Survey on techniques related to Sentiment analysis of Twitter posts

Rohit Shukla¹, Nishchol Mishra² and Sanjeev Sharma³

¹P.G. student School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Airport Bypass Road, Gandhi Nagar, Bhopal, Madhya Pradesh India
²Associate Professor School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Airport Bypass Road, Gandhi Nagar, Bhopal, Madhya Pradesh India
³Associate Professor School of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Airport Bypass Road, Gandhi Nagar, Bhopal, Madhya Pradesh India

Abstract

Sentiment analysis is the process of collecting views of the peoples about any topic, subjects or any item. To analyze views NLP (natural language processing) and linguistic based techniques are used. This paper presents survey over the various techniques which are used for sentiment analysis. For this purpose, data of social media sites like twitter are used to collect various views about any product or item. There are techniques like LDA (Latent Dirichlet Allocation), are used to predict views about the data.

Keywords: - Sentiment analysis, NLP (Natural Language Processing), LDA (Latent Dirichlet Allocation).

Introduction

Sentiment analysis is a type of natural language processing for finding the state of mind of the public about a particular topic or product. Sentiment analysis, that is additionally known as opinion mining, involves in building a system to collect and examine the state of mind or opinions about the product made in the blog post, reviews, comments or tweets. There are lots of challenges in sentiment Analysis. The first is an opinion word that’s thought of to be positive in one scenario is also thought of negative in another scenario. A second challenge is that individuals don’t continuously express opinions within the same manner. Most traditional text process depends on the very fact that little variations between two items of text don’t amendment the that means pretty much.

Sentiment analysis concentrates on attitudes, whereas text mining target the analysis of reality. There are few main fields of analysis predominate in sentiment analysis: sentiment classification [1], feature based sentiment classification[2] and opinion summarization. Sentiment classification deals with categories entire documents per the opinions towards sure objects. Feature-based sentiment classification[2] on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions.

Twitter (www.Twitter.com) is a colossal social networking site for fast communication.140-word 500 million "tweets" are posted by more than 316 million monthly active users each day. In the India's election ʹͲʹͲ Twitter and other social networking sites play big role in winning of Mr. Narendra Modi. 80% Active user are using twitter from mobile
and makes Twitter so big in developing country like India where 117 million Smartphone users. Because micro blogging websites become a source of various kinds of information.

Twitter's quality as data supply has led to the event of application and analysis in varied domain. Humanitarian help and Disaster Relief is one domain wherever data from twitter is employed to supply situational awareness to a crisis[3].for most people throughout the decision-making method. Long before awareness of the world Wide internet became widespread, many individuals asked our friends to advocate a self-propelled vehicle mechanic or to elucidate who they were planning to vote for in native elections, requested reference letters regarding job candidates from colleagues, or consulted client Reports to come back to a choice what dishwasher to buy for. however, the web and also the net have currently (among alternative things) created it doable to search out concerning the opinions and experiences of these within the Brobdingnag pool of individuals that area unit neither our personal acquaintances nor well-known skilled critics — that's, individuals we’ve never detected of. And conversely, a lot of individual’s area unit creating their opinions out there to strangers via the web.

Due to the sheer volume of opinion made internet resources like discussion forum, review sites, blogs and news corpora out there in digital from, a lot of the present analysis is specializing in the realm of sentiment analysis. folks are supposed to develop a system that may establish and classify opinion or sentiment as diagrammatic in a textual matter.

Collecting twitter data

500 million tweets are generated everyday some of these are available to researchers and practitioners through public APIs at no cost. APIs to get to Twitter information can be characterized into two sorts in view of their configuration and access strategy:

REST APIs depend on the REST structural planning now prevalently utilized for outlining web APIs. These APIs utilize the draw system for information recovery. To gather data a client should expressly ask for it. Streaming APIs gives a ceaseless stream of open data from Twitter. These API utilizes the push methodology for information recovery. Once a solicitation for data is made the spilling APIs give a constant stream of overhauls with no further information from the client.

They have diverse capacities and confinement as for what and the amount of data can be recovered. The gushing API has three sorts of endpoints:

- Public Streams: Public tweets are streamed from Twitter.
- User Streams: Users see tweets by particular user.
- Site streams: Twitter provide site stream so that application can access tweets from multiple users.

The public streams API is the most versatile streaming API. Requests to the APIs contain parameters which can includes hashtags, keywords, geographic regions and Twitter user IDs. Responses from Twitter API is in JavaScript Object Notion (JSON) format. JSON is a popular format that is widely used as an object notation on the web.[4]
collection a user’s Tweets

Status messages are called as tweets on twitter. Length of tweet can be at most 140 characters. Desktop and mobile clients are used to publish tweets through the twitter API. When one user post tweets which is created by another user that tweet is called as retweet. Rest and streaming API can be used to access the user’s tweet.

We can access a user’s tweets by using $\text{statuses/user\_timeline}$ from the REST APIs. Using this API, can retrieve 3,200 of the most recent Tweets published by a user including retweets. The API returns Twitter “Tweet” object. Specifically, the $\text{statuses/filter}$ API[5] provides a constant stream of public Tweets Published by user. We can create a POST request to the API and fetch the search results as a stream.

Searching on Twitter is facilitated through the use of parameters. Acceptable parameter values for search include keywords, hashtags, phrases, geographic regions, and usernames or userids. Twitter search is quite powerful and is accessible by both the REST and the Streaming APIs. There are certain subtle differences when using each API to retrieve search results.

Twitter provides the $\text{search/tweets}$ API[5] to facilitate searching the Tweets. The search API takes words as queries and multiple queries can be combined as a comma separated list. Tweets from the previous 10 days can be searched using this API.

Using the Streaming API, we can search for keywords, hashtags, userids, and geographic bounding boxes simultaneously. The filter API facilitates this search and provides a continuous stream of Tweets matching the search criteria. POST method is preferred while creating this request because when using the GWT method to retrieve the results, long URLs might be truncated.

Obtaining data via Reseller

The rate limitations of Twitter APIs can be too restrictive for certain types of applications. To satisfy such requirements, Twitter Firehose provides access to 100% of the public Tweets on Twitter at a price. Firehose data can be purchased through third party resellers of twitter data.

Twitter data some of them also provide data from other social media platforms, which might be useful while building social media based systems. These include the following:

- DataSift - provides access to past data as well as streaming data
- GNIP – provides access to streaming data only
- Topsy – provides access to past data only.

analyzing twitter data

The data we collect from Twitter quickly grows to immense proportions. In fact, they grow so large that attempting to read each individual Tweet quickly become a hopeless cause. A more reachable goal is to get a high-level understanding of what our users are talking about. One way to do this is by understanding the topics the users are discussing in their Tweets. The automatic discovery of topics in the text through “topic modeling” with latent Dirichlet allocation (LDA), a popular topic modeling algorithm.
Every topic in LDA is a collection of words. Each topic contains all of the words in the corpus with a probability of the word belonging to that topic. So, while all of the words in the topic are the same, the weight they are given differs between topics.

LDA find the most probable words for a topic, associating each topic with a theme is left to the user.

**LITERATURE REVIEW**

[6] the authors worked on POS-specific prior polarity features and use of a tree kernel to obviate the need for tedious feature engineering. In this paper, they look at one such popular microblog called Twitter what’s more, form models for grouping "tweets" into positive, negative and nonpartisan sentiment. They fabricate models for two characterization assignments: a paired undertaking of arranging conclusion into positive and negative classes and a 3-route errand of grouping feeling into positive, negative and impartial classes. They experiment with three types of models: unigram model based model, a tree kernel, and a feature based model.

They investigated two kinds of models: tree kernel and feature based models and demonstrate that both these models outperform the unigram baseline. For our feature-based approach, they do feature analysis which reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags. They tentatively conclude that sentiment analysis for Twitter data is not that different from sentiment analysis for other genres.

In [7] author propose a supervised sentiment classification framework that relies on information from Twitter, a preferred microblogging service. By utilizing 50 Twitter tags and 15 emoticons as sentiment labels, this framework avoids the requirement for effortful manual annotation, permitting identification and classification of various sentiment kinds of short texts. the standard of the sentiment identification was additionally confirmed by human judges. They additionally explore dependencies and overlap between completely different sentiment varieties painted by smileys and Twitter hashtags.

In their study, they use four completely different feature varieties (punctuation, words, n-grams and patterns) for sentiment classification and value the contribution of every feature sort for this task. They evaluated numerous feature varieties for sentiment extraction together with punctuation, patterns, words and n-grams, confirming that every feature sort contributes to the sentiment classification framework. They also proposed two different methods which allow an automatic identification of sentiment type overlap and inter-dependencies. In the future these methods can be used for automated clustering of sentiment types and sentiment dependency rules. While hashtag labels are specific to twitter data, the obtained feature vectors are not heavily Twitter-specific and in the future they would like to explore the applicability of Twitter data for sentiment multi-class identification and classification in other domains.

In [8] the author tries find user interest in microblogs by novel topic model mining model called as User Topic Model (UTM). This model is based on Latent Dirichlet Allocation. Author divided the interests in two parts by way of generation in the microblogs: original interest and retweet interest. So the model can discover user's original interest and retweet interest than mine the user's interest words. The experimental results shows that UTM outperform the LDA and twitter-user model[9] for discovering user interest.
Methodology used for dig out user interest by UTM: The LDA model is a generative process where each document in the text corpus is modeled as a set of draws from a mixture distribution over a set of hidden topics. A topic is modeled as a probability distribution over words. The Gibbs sample is used to infer latent variable of topic space.

In [10] a description of Social media sites is presented. Social media sites such as Twitter and Facebook play an important role in students' life to communicate and exchange information. This paper focuses on student data generated on social media sites rather than the academic information. For that they uses social media tool Radian6 to collect the data from Twitter about students' college experiences.

[11] previous analysis principally targeted on modeling and following public sentiment. during this work, they move one step more to interpret sentiment variations, they ascertained that rising topics at intervals the sentiment variations amount is extremely associated with the real reasons behind the variation. supported this observation they propose a latent Dirichelet Allocation based mostly model, foreground and background LDA(FB-LDA), to distill foreground topic and filter out longstanding background topics. These foreground ground topics will provide potential interpretations of the sentiment variations. To more enhance the readability of the deep-mined reasons, they choose the foremost representative tweets for foreground topics and develop another. The RCB-LDA model can rank a set of season candidates expressed in natural language to provide sentence-level reasons. Their proposed models were evaluated on real Twitter data.

In [12] author presents SentiView, an intuitive representation framework that intends to examine open assumptions for famous themes on the web. SentiView joins instability successive words in the content information, it mines and models the progressions of the sentiment on open subjects.

This paper has introduced a new visualization system for analyzing visualizing and verifying the sentiment mining method and a model-driven prediction approach have been used to analyze the public sentiment on hot topic. Considering the characters and interests among different participants in some popular topic. They have proposed two new visualization concepts, the helix combined with astrolabe and relationship map to visualize change of multiple attributes and complex relations among the attributes such as numbers, location distribution, ages and sentiments of the participants.

**CONCLUSION**

While Facebook played a very important role at creating groups, organizing and mobilizing at the initial stages, Twitter helped the revolution gain momentum by recording events as they happen and spreading news to the world. The sentiment carried by Facebook or Twitter posts definitely inspired and galvanized people for more action. A careful review of the literature on sentiment analysis showed that there is no one best feature vector that is suited to sentiment analysis. There are some sentiment analysis studies that achieved good results with unigram presence and other studies with bi-gram presence. Also there are sentiment analysis studies which indicated that incorporating syntax, negation and/or POS tags improved performance of sentiment analysis.
References


INTRODUCTION

Today, success of electronic market depends on highest level of system integration in order to achieve low power and low cost without compromising speed or performance for mass production in the consumer, wireless and computer market. System-on-Chip (SoC) allows to integrate analog as well as digital circuits on a single chip in order to achieve analog mixed signal devices such as ADC, PLL, and DAC etc. In nanoscale CMOS technologies, due to feature size reduction it is necessary to scale down supply voltage to achieve reliable operation of transistors. Phase locked loop (PLL) is one of the most widely used mixed signal device used for various applications such as on-chip clock generations in various applications such as frequency synthesizers, frequency tracker and clock recovery circuits. But PLL is one the most power consuming device and each and every block should be optimized in order to minimize overall power consumption and to make design power efficient.

1.1 PHASE LOCKED LOOP

Phase Locked Loop is basically a closed loop frequency control system. PLL Clocks are used when the system needs to minimize the propagation delay. It is able to do this by acting as a phase detector to keep an input clock in phase with an output frequency of VCO which is fed back to phase detector through feedback network. As the Fig-1 suggests, Functional building blocks of PLL are as follows:

1. Phase Frequency Detector
2. Loop Filter
3. VCO

Fig -1: Block Diagram of Traditional PLL

Phase Frequency Detector compares the feedback frequency from VCO and input reference frequency and generates a DC level which is proportional to Δφ (i.e. Phase Difference of input and output) [1]. Low Pass Filter as the name suggests filter out high frequency components and allow DC components to achieve pure DC V\text{control}. It is an oscillator which is controlled by a DC voltage and generate output frequency based on it. Though this type of PLL is comprised of mostly analog blocks and it is difficult to design analog circuits in deep submicron CMOS technology at low supply voltage. As a result poorly performed PLL can be a bottleneck in the current and coming computational systems.

1.2 ADVANCES IN PLL

In SoCs, there are multiple functional blocks which require different clock frequencies thus clock generators with wide tuning-range are essential. In low voltage applications, wide frequency range may cause a large VCO gain which can increase VCO sensitivity towards power/ground noises. Earlier, LC-VCOs were used to achieve smaller output jitter and good phase noise performance with low power consumption. However, the tuning-range of LC-VCO (around 10-20%) is relatively less compared to RO-VCO (>50%) [5]. As a result output frequency may fall out of desired range due to process variations.

Also phase noise performance of PLL depends on quality factor of inductor and it is difficult to achieve required quality factor in digital CMOS technology. As Spiral inductor in LC-VCO occupies a large amount of chip area which reduces advances to be achieved by scaling [6]. At low supply voltages large threshold voltages of transistors may cause CMOS switching to slow down which can prevent VCO and frequency divider to operate on high frequencies. There are several methods developed to design a low voltage analog circuit design: (1) special low-VT devices [10], (2) on-chip clock and gate voltage boosting [11] and (3) body-biasing and body-driven circuits [12] [13]. Low-VT devices requires extra masks during fabrication and incurs higher production costs. On-chip voltage boosting introduces long-term reliability concerns, especially for nanoscale CMOS devices. Body-biasing and body-driven circuits cause latch-up problems. To avoid forward biasing of the body–source junction, the PMOS body (NMOS) connects to the VDD (GND) [13].

A 90-350 MHz LVPLL has been proposed using 0.5V supply voltage. Section 2 shows architecture of proposed PLL. Conclusions are drawn in Section 3.

2. ARCHITECTURE OF PROPOSED LVPLL

Architecture of proposed LVPLL is shown in Fig-3. Here, the MCSS charge pump (CP) circuit operates at low supply voltage in order to achieve higher supply voltages to drive analog circuit blocks and its small parasitic capacitance helps to decrease switching time of PMOS and NMOS resulting in reduction in leakage current. The Low voltage VCO is a Ring oscillator (RO) which consists of 4-stage delay cells and a low voltage segmented current mirror (LV-SCM) for biasing. To operate over wide frequency range, low K\text{VCO} is essential thus LV-SCM is added with LV-VCO to achieve a low K\text{VCO} and multi-band frequency output. Frequency Divider is used to enhance operating speed of LVPLL.
2.1 MCSS CHARGE PUMP

Various Charge Pump Architectures have been reported serving different purposes to different targeted applications [2]. Single Ended CPs are popular in SoCs as they do not require any additional loop filters and consume low power operating at low supply voltage. In the previous work, switches at the drain of the current mirror causes charge injection in DC control voltage of VCO \( V_{cntrl} \) which make PLL not to remain in locked condition as VCO is not able to maintain a minimum constant phase difference between its output frequency and input frequency [14]. Furthermore, there is a charge sharing problem between loop filter capacitor and MOS parasitic capacitance. The cascade switches in Single ended Charge Pump requires voltage headroom for given supply voltages \( V_{DD} > V_{DSP} + V_{DSN} + V_{DSATP} + V_{DSATN} \) where \( V_{DSP}, V_{DSN} \) are drain to source voltages of PMOS and NMOS respectively and \( V_{DSATP}, V_{DSATN} \) are drain source saturation voltages of PMOS and NMOS respectively. Thus, these previous designs of Charge pump are facing current mismatch during switching activities and charge sharing in VCO control voltages thus these designs are not suitable for low voltage operations [7].

A Modified Charge Sharing Scheme (MCSS) is proposed here. In this scheme, CP switches control gate instead of drain or source which creates a current path from power to ground and dc leakage current \( I_L \) of M1 transistor causes a voltage drop down in \( V_{cntrl} \). This leakage current can be reduced by inserting two switches SWUP and SWDN to control UP and DN signals respectively. As shown in Fig-3, when UP=0.5 and DN=0 SW2 and SWUP are turned on and SW1 and SWDN are turned off. With this arrangement current path is cut off and it will minimize leakage current. In addition, M3 an M4 are remains in saturation region thus circuit produces less noise and higher operational switching activities compared to conventional CP.

2.2 LV-VCO

VCO is an important component of PLL and it plays an important role to achieve wide tuning-range at low supply voltage and to achieve a power efficient system. Conventional ring VCOs have some advantages to offer [5]. To obtain high output frequency, LC-VCO is a good choice as it provides high noise rejection capability but it occupies large area than ring oscillator on chip and faces difficulty to produce multi-phase output frequency [3]. In the proposed LV-VCO, 4 stage differential ring oscillator is chosen to reduce supply noise and to produce 8-phase output frequency. Fig-4 shows LV-VCO consist of 4-stage delay cells and LV-SCM. LV-SCM converts digital code into analog current so delay cells can produce multi band output frequencies. For low voltage applications, cascaded design of delay cell must be avoided in order to prevent supply noise in output frequency [4]. Drain current of PMOS transistor is:

\[
I_D = \frac{1}{2} \mu C_{ox} \frac{W}{L} (V_{SG} - |V_{TP}|)^2 (1 + \lambda V_{SD})
\]  

(1)

The delay cells of LV-VCO consists of a pair of differential input NMOS and a pair of PMOS load. Current through PMOS load can be adjusted by tuning the \( V_{cntrl} \) which gives a small \( KVCO \) compared to conventional PMOS loads which leads to a wider range of output frequency compared and produces multi-phase signals.

![Fig -3: Schematic of MSCC Charge Pump](image1)

![Fig -4: 4-stage Delay Cells in RO](image2)

![Fig -5: Layout of LV-Ring VCO in Microwind](image3)
2.3 FREQUENCY DIVIDER

Frequency Divider is an essential building block of PLL. As the output frequency of VCO might turn out in MHz or GHz range, it is very difficult to get a precise phase difference between input and output frequencies. Due to high input frequency, FD cannot be implemented in conventional static CMOS logic. Dynamic flip-flops are faster and more compact compared to static ones. In proposed work, a common dynamic flip flop variety True Single Phase Clock (TSPC) is used which performs flip-flop operation with high speed and low power [8]. The architectures of TSPC are based on edge triggering phenomena i.e. Positive and Negative. Among these two architectures, negative edge triggered architecture for dynamic latches are popular for high speed circuit design. In order to obtain sharp clock signals, NMOS devices are used as a stronger pull-down in negative edge triggered CMOS devices. In FD, first divider stage has maximum operating speed as FD has a high input frequency. As per the requirement, input frequency is divided by a fixed factor \( N \). In this proposed work, divide by 2 network is made up of TSPC D-FF. This proposed FD reduces the power consumption and provide higher operating speed.

3. CONCLUSION

In this paper, 90-350 MHz Phase Locked Loop operating at 0.5V supply voltage is implemented in standard 90nm CMOS technology. The proposed Charge Pump reduces the leakage current by 60% and increases switching speed. Low voltage RO-VCO achieves high speed, wide tuning range and low KVCO gain. Proposed PLL can be used for most low power applications and can be used in wireless communication.
ACKNOWLEDGEMENT
The authors would like to thank Microwind EDA solutions and Mr. Vinay Sharma (ni2 designs, Pune-India) for providing licensing support by which the designs, simulations and layouts used in this paper were possible.

REFERENCES